Rice University

Computational Biology: Insights into Hemagglutinin and Polycomb Repressive Complex 2 Function

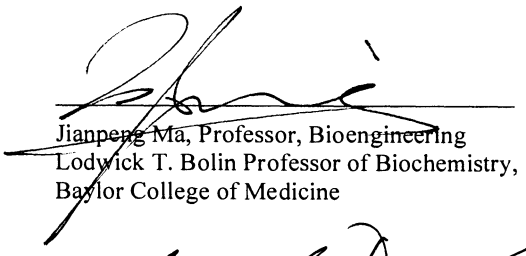
By

Brian David Kirk

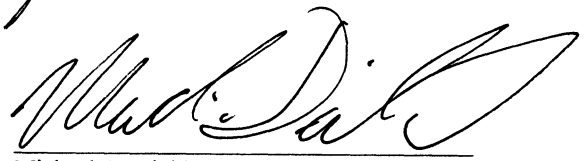
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

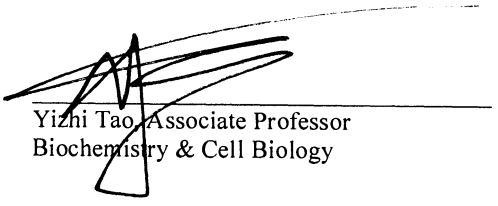
APPROVED, THESIS COMMITTEE:



Jianpeng Ma, Professor, Bioengineering
Lodwick T. Bolin Professor of Biochemistry,
Baylor College of Medicine



Michael R. Diehl, Assistant Professor
Bioengineering, Chemistry



Yizhi Tao, Associate Professor
Biochemistry & Cell Biology

Houston, Texas
May, 2012

ABSTRACT

Computational Biology: Insights into Hemagglutinin and Polycomb Repressive Complex 2 Function

By

Brian D. Kirk

Influenza B virus hemagglutinin (HA) is a major surface glycoprotein with frequent amino-acid substitutions. However, the roles of antibody selection in the amino-acid substitutions of HA were still poorly understood. An analysis was conducted on a total of 271 HA_1 sequences of influenza B virus strains isolated during 1940~2007 finding positively selected sites all located in the four major epitopes (120-loop, 150-loop, 160-loop and 190-helix) supporting a predominant role of antibody selection in HA evolution. Of particular significance is the involvement of the 120-loop in positive selection. Influenza B virus HA continues to evolve into new sublineages, within which the four major epitopes were targeted selectively in positive selection. Thus, any newly emerging strains need to be placed in the context of their evolutionary history in order to understand and predict their epidemic potential.

As key epigenetic regulators, polycomb group (PcG) proteins are responsible for the control of cell proliferation and differentiation as well as stem cell pluripotency and self-renewal. To facilitate experimental identification of PcG target genes, which are poorly understood, we propose a novel computational method, *EpiPredictor*, which models transcription factor interaction using a non-linear kernel. The resulting targets suggests that multiple transcription factor networking at the *cis*-regulatory elements is critical for PcG recruitment, while high GC content and high conservation level are also important features of PcG target genes.

To try to translate the *EpiPredictor* into human data, we performed a computational study utilizing 22 human genome-wide CHIP data to identify DNA motifs and genome features that would potentially specify PRC2 using five motif discovery algorithms, Jaspar known transcription binding motifs, and other whole genome data. We have found multiple motifs within the various subgroups of experimental categories that have much higher enrichment against CHIP identified gene promoter than among random gene promoters. Specifically, we have identified Low CpG content CpG Islands (LcG's) as being critical in the separation of Cancer cell line identified targets from Embryonic Stem cell line identified targets. Additionally, there are differences between human and mouse ES cell predictions using the same motifs and features suggesting relevant evolutionary divergence.

ACKNOWLEDGEMENTS

I would like to thank the members of my Thesis Committee, Drs' Jianpeng Ma, Michael R. Diehl, and Yizhi Tao; Dr. Qinghua Wang for the years of close mentoring she has provided; and my lab mates and collaborators: Jun Shen, Yufeng Gou, and Jia Zeng, without whom none of this would have been possible.

I am thankful to Drs. Leonie Ringrose and Marc Remsmeier for providing valuable details regarding their publications. We also thank Kit Menlove and Darren Seibert for their input on the project. The work was in part supported by the Collaborative Advances in Biomedical Computing from The John and Ann Doerr Fund for Computational Biomedicine at Rice University. Funding in part has come from the NLM Training Program of the Keck Center of the Gulf Coast Consortia (NIH Grant No. 5 T15LM07093-17).

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

Chapter 1. Introduction

We live in a world of data. There are over 21 million citations for biomedical literature in PubMed which represents about 6.3 billion words worth of information, and that is just from abstracts[1]. There is also the experimental data itself, already, the complete genome sequences from more than 180 different organisms have been mapped[2]; or newer whole genome expression (RNA-seq) or mapping experiments (CHIP-seq or Hi-C) which yield lists of potentially several thousands of genes from each experiment. Each set representing that which could be important in a given biological process or cell type. Then there is the fact that there are at least 210 different cell types, not including cancerous or other diseased cell phenotypes, in the human body, each with its own unique cellular programs responsible for its maintenance and function[3]. All of this is just to point out that there is so much data already generated that it becomes critical to contextualize that which already exists often before new insights can be made.

As such, bioinformatics and data analysis have become increasingly necessary tools as whole genome experiments have become more common. Fortunately, the complexity and quantity of data has been accompanied by parallel increases in computational power. Consequently, the modern scientist has at their disposal a host of sophisticated tools and techniques from which to attempt to wring useful information and knowledge from the vast expanse of available data[4].

One of the most potentially useful of the techniques is that of the Machine Learning Classifier (MLC). This type of analysis allows for both the mining of data

for previously unknown connections and the prediction of classification state based on said relationships. MLC's can have a large variety of forms and algorithms which are divided according to their purpose. These being primarily supervised, or unsupervised learning, which differ from each other in that supervised learning uses some sort of known input in order to classify a much larger unknown dataset, whereas unsupervised learning only attempt to classify data into its distinguishing components, a very common application of this being clustering programs[4]. There are also many other varieties that in some way combine supervised and unsupervised principles in their function. The importance of these techniques is that it allows data, which can exist in complicated multidimensional spaces, to be sorted, grouped, and analyzed according to the things it is most similar to according to the functional outcome of a given experiment, and to thus make predictions about what else should then share that functional outcome in a different experiment, or maybe even predict which data will share a completely different functional outcome from a different test altogether[5].

MLC's are so powerful in a biological context because they don't care what type of data that you are presenting them. The data in question could be lists of genes, DNA or amino acid sequences, or physical measurements; such as expression data or fluorescence[6]. Particular to my interests, are the applications of these algorithms to phylogenetic sequence organization[7], transcription factor binding prediction[8], and epigenetic state prediction and as such, the main goal of the completed work presented here is to illustrate how MLC's can be applied in the analysis and subsequent insight into the biological function of two important

biological systems: Influenza B virus hemagglutinin (HA) evolution and Polycomb Repressive Complex 2 (PRC2) protein targeting.

To do such, requires that a brief introduction into the theory behind one of the MLC's that will be presented, Support Vector Machine (SVM). SVM is a supervised learning technique which seeks to separate data in higher dimensional space through the use of a non-linear hyperplane. The optimization of SVM is then to determine the hyperplane which is able to maximize the distance of the classified training data to the hyperplane, and then apply that separating hyperplane to a larger set of data[9]. A simple 2D illustration of SVM methodology is shown in Figure 1.1.

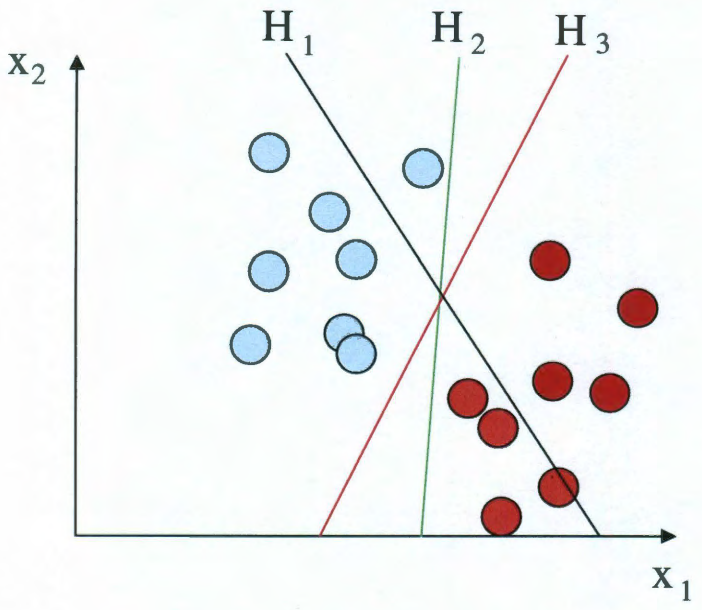


Fig. 1.1. **2D schematic illustration of hyperplane separating data.**

Let this 2D data set represent a much higher order series of data. In it a series of training data has been labeled by the functional classification that has been

observed in experimental results, red for one result and blue for another. It is then possible to separate the data mathematically with the application of a hyperplane. Both H_2 and H_3 represent hyperplanes capable of separating the data, but H_3 yields the maximum distance of separation and is thus established as the be classifier. Additionally, it possible to do this in transformed space using non-linear kernel functions which allows for a linear hyperplane in the transformed space to allow for nonlinear classification in real space.

Chapter 2. Influenza B Virus Hemagglutinin

2.1 Background

Ever since the isolation of the first influenza B virus strain B/Lee/40 [10], influenza B virus has remained a serious health problem, contributing to the seasonal "flu" epidemics each year. As a major glycoprotein on the surface of influenza B virus, HA undergoes constant amino-acid substitutions. The HA protein of current circulating influenza B virus strains belongs to one of the two major phylogenetic lineages: B/Victoria/2/87 (B/VI)-like and B/Yamagata /16/88 (B/YM)-like [11-13].

Over the last 68 years, a large number of amino-acid substitutions on influenza B virus HA were observed in field isolates, in monoclonal-antibody escape mutants and in egg-adapted variants [10-11, 14-22]. However, it was unclear which of these substitutions were the results of positive selection, and what were the roles of antibody selection in the molecular evolution of influenza B virus HA. In this context, **positive selection** is defined as a significant excess of amino-acid altering substitutions over silent substitutions in nucleotide sequences, since, if completely random, only 24% of nucleotide substitutions would cause changes in the encoded amino acids [23].

There was previously a sequence analysis on 49 HA_1 sequences of recent influenza B virus isolates, which identified HA_1 75, 197, and 199 (B/HongKong/8/73 HA numbering of 75, 194, and 196, respectively) to be under positive evolutionary selection [24]. However, since it did not separate the B/YM-lineage and B/VI-lineage strains, this study might have failed to identify those amino-acid positions that were

selected positively only in one but not the other lineage [25]. This was particularly problematic for the 150-loop and 160-loop (**Fig.2.1**), which had become specific for B/YM-like and B/VI-like strains, respectively [26-30]. Most recently, a larger-scale analysis that used 214 HA_1 sequences of influenza B virus strains has been published [31]. Although it separated B/YM- and B/VI-lineage strains, the evolution of these lineages into distinct sublineages was not taken into account, which limited the accuracy of the positively selected sites derived therein [31].

PAML is a package of programs that analyze DNA and protein sequences using maximum likelihood [32]. Using the program CODEML in PAML, the nonsynonymous (amino-acid altering substitutions)/synonymous (silent substitutions) rate ratio (ω) for each codon is calculated as an important indicator of selection pressure at the protein level: an $\omega > 1$ indicates positive selection [25, 32]. Bayes Empirical Bayes analysis then calculates the posterior probability that each site belongs to a particular site class. Sites with high posterior probability of belonging to the site class of $\omega > 1$ are inferred to be under positive selection [25, 32].

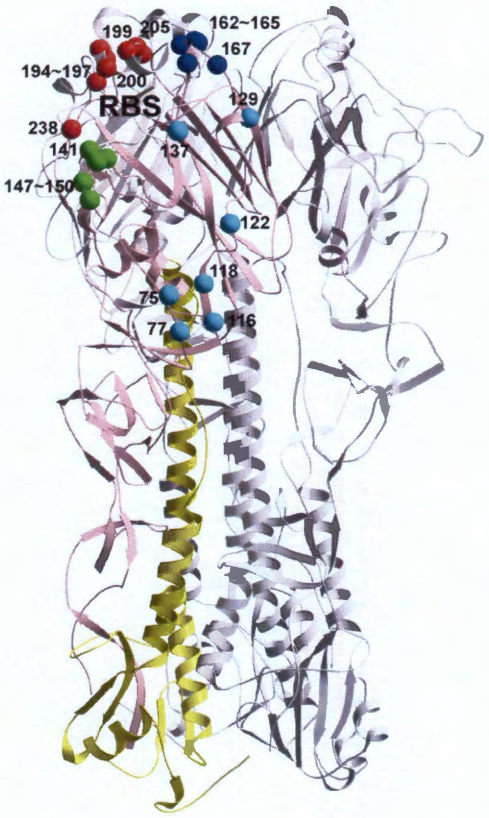


Fig. 2.1. Major epitopes of influenza B virus HA.
The trimeric HA is shown with one monomer highlighted in color: pink for HA_1 and yellow for HA_2. Mutations in four regions, the 120-loop (cyan), 150-loop (green), 160-loop (blue), and 190-helix (red), have been found to cause antigenicity variation. The receptor-binding site (RBS) is labeled.

In order to gain insights into the amino-acid positions on influenza B virus HA that are truly under positive selective pressure, here a total of 271 HA_1 sequences of influenza B virus strains isolated between 1940~2007 were analyzed. Based on the phylogenetic analysis, these HA_1 sequences were divided into three major groups: early strains (1940~1970), B/YM-like lineage (1972~2005) and B/VI-like lineage (1975~2007). The B/YM-lineage was further divided into four sublineages, and the B/VI lineage into two sublineages (**Fig.2.2**). These seven groups were analyzed by using CODEML in PAML version 4 [25, 32]. The identified positively selected sites

were located predominantly on the four major antigenic epitopes on HA_1: the 120-loop (HA_1 116~137), the 150-loop (HA_1 141~150), the 160-loop (HA_1 162~167), the 190-helix (HA_1 194~202), and their respective surrounding regions [30] (**Fig.2.3**), suggesting the important roles of antibody selection in molecular evolution of influenza B virus HA.

2.2 Materials and Methods

Phylogenetic analysis

This study focused on the first 340 amino acid residues of mature HA_1 (1~1020 nucleotides excluding those corresponding to the signal peptide). A total of 271 HA_1 sequences of influenza B virus strains isolated between 1940~2007 were used in the study. These sequences were selected to sample all the years in which influenza B viruses were active, and special cares were taken to avoid very similar stains isolated in the same regions. All the sequences were obtained from the Influenza Sequence Database (Los Alamos National Laboratory, Los Alamos, NM, USA www.flu.lanl.gov) [33]. The CLUSTAL W method [34] with the MEGALIGN program of DNASTAR package (www.dnastar.com) was employed for sequencing alignment and phylogenetic analysis.

Analysis of selective pressure

For this analysis, the CODEML program in PAML was used to calculate the codon-substitution models for heterogeneous selection pressure at amino-acid positions [7, 25, 32, 35]. The models used in this study were M0, M1a, M2a, M7 and

M8. M1a (nearly neutral) and M7 (beta) were null models that did not support $\omega > 1$. In contrast, the alternative models M2a (positive selection) and M8 (beta and ω), compared to M1a and M7 respectively, each had an additional class that allowed $\omega > 1$. Likelihood ratio tests (LRT) comparing M2a versus M1a and M8 versus M7 provided test for the existence of positive selection. In LRT, twice the log likelihood difference, $2\Delta l = 2(l_1 - l_0)$, was compared with a χ^2 distribution to test whether the null model was to be rejected, where ℓ_1 and ℓ_0 were the log likelihood for the alternative model and the null model, respectively. In addition, empirical Bayes analysis was employed to calculate the posterior probability that each site belonged to a particular site class. Sites with high posterior probability of belonging to the site class of $\omega > 1$ were inferred to be under positive selection. It was shown that Bayes Empirical Bayes, which assigned a prior to the model parameters [36], worked well for both small and large datasets [35]. Since some of the subgroups used in this study were small, the results from Bayes Empirical Bayes analysis were used throughout this study. To account for the insertions and deletions in influenza B virus HA_1, the numbering of influenza B/HongKong/8/73 HA was used as a reference for all sequences [30].

2.3 Results

Phylogenetic relationship of influenza B virus HA

According to phylogenetic analysis, the 271 HA_1 sequences were divided into three groups: early strains isolated between 1940~1970 (I), B/YM-like lineage since 1972 (II) and B/VI-like lineage since 1975 (III). The B/YM-lineage (II) was divided further into four (II-*i* ~ II-*iv*) sublineages (**Fig.2.2**), among which (II-*i* ~ II-*iii*) sublineages had been described in a previous study [37], whilst the (II-*iv*) sublineage was described here for the first time. The B/VI-lineage (III) was divided further into an earlier sublineage (III-*i*) and a more recent sublineage (III-*ii*) (**Fig.2.2**). This large-scale phylogenetic analysis uncovered that the divergence of influenza B virus HA into B/YM- and B/VI-lineages can be dated back to early 1970s, which is much earlier than previously thought [11, 13, 38-39] and agrees well with a just-published study [40].

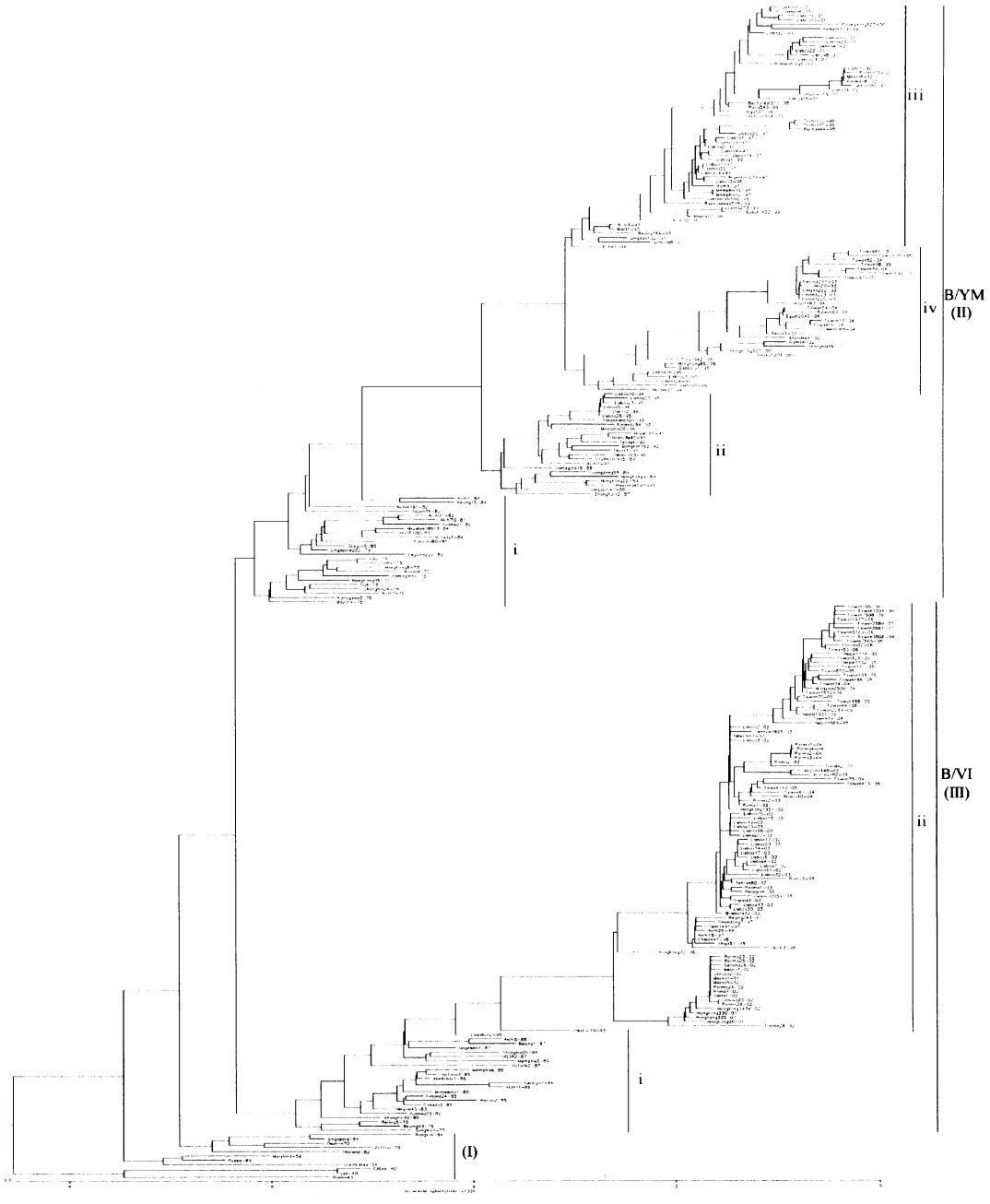


Fig.2.2. Phylogenetic relationship of 271 HA_1 sequences used in this study.
For all sequences, the nucleotide sequences between 1~1020, corresponding to residues HA_1 1~340, were used. The phylogenetic tree was drawn using the program Megalign from DNASTAR package (www.dnastar.com)

Positive selection on influenza B virus HA

To detect positively selected sites in HA_1 sequence of influenza B virus strains between 1940~2007, the analysis using CODEML in PAML was performed individually on the seven subgroups (I, II-*i* ~ II-*iv*, and III-*i* ~ *ii*) (**Fig.2.2**). In all but two cases, the LRT statistics ($2\Delta l$) for M2a versus M1a and M8 versus M7 were much larger than the critical value of $\chi^2_{1\%} = 6.63$ with degree of freedom (d.f.) set to 1 (**Tables 2.1, 2.2**). Thus the LRT tests supported the existence of positive selection on influenza B virus HA. The sites with greater than 50% posterior probability to be under positive selective pressure in models M2a and M8, obtained from Bayes Empirical Bayes analysis [35], were listed in **Table 2.3**. In general, M2a identified fewer sites under positive selection than M8 did. Nevertheless, the sites identified in M2a were those of the highest posterior probability in M8 (**Table 2.3**). In contrast, those identified only in M8 but not in M2a were generally of low posterior probability. To be more conservative, most of our discussion was focused on the sites that were identified in M8 model with greater than 95% posterior probability to be under positive selection. This cutoff limits the false-positive rate to 5~6% or lower [35]. It is important to emphasize that those of high posterior probability to be under positive selection were not necessarily those of the highest mutation rates. Different from influenza A virus HA [25, 41], a much smaller number of sites on influenza B virus HA were subject to positive selection for antigenic drift, consistent with earlier studies [23, 31].

Table 2.1. The values of log-likelihood (ℓ), d_N/d_S, and parameter estimates in the analysis of the HA_1 subunit of influenza B virus strains circulating between 1940~2007

Model	ℓ	d_N/d_S	Parameters estimates
Early strain (I) 1940~1970 (11 strains)			
M0 (one-ratio)	-2343.24	0.271	ω=0.271
M1a (nearly neutral)	-2317.53	0.258	p_0=0.744 (p_1=0.256), ω_0=0.002 (ω_1=1)
M2a (positive selection)	-2314.85	0.297	p_0=0.744, p_1=0.251 (p_2=0.005), ω_0=0.005 (ω_1=1), ω_2=7.990
M7 (beta)	-2317.89	0.236	p=0.016, q=0.051
M8 (beta&ω)	-2315.01	0.287	p_0=0.994 (p_1=0.006), p=0.017, q=0.051, ω_s=7.428
B/YM-lineage (II-*i*) 1972~1984 (25 strains)			
M0 (one-ratio)	-2501.04	0.373	ω=0.373
M1a (nearly neutral)	-2461.96	0.262	p_0=0.755 (p_1=0.245), ω_0=0.022 (ω_1=1)
M2a (positive selection)	-2439.17	0.404	p_0=0.729, p_1=0.261 (p_2=0.011), ω_0=0.020 (ω_1=1), ω_2=11.904
M7 (beta)	-2462.35	0.300	p=0.005, q=0.012
M8 (beta&ω)	-2439.28	0.426	p_0=0.989 (p_1=0.011), p=0.017, q=0.042, ω_s=12.282
B/YM-lineage (II-*ii*) 1987~1996 (24 strains)			
M0 (one-ratio)	-2032.98	0.311	ω=0.311
M1a (nearly neutral)	-2002.40	0.188	p_0=0.812 (p_1=0.188), ω_0=0 (ω_1=1)
M2a (positive selection)	-1995.15	0.318	p_0=0.826, p_1=0.136 (p_2=0.038), ω_0=0 (ω_1=1), ω_2=4.779
M7 (beta)	-2002.45	0.200	p=0.005, q=0.020
M8 (beta&ω)	-1995.32	0.313	p_0=0.953 (p_1=0.047), p=0.012, q=0.076, ω_s=4.237
B/YM-lineage (II-*iii*) 1991~2002 (56 strains)			
M0 (one-ratio)	-2211.56	0.200	ω=0.200
M1a (nearly neutral)	-2196.38	0.168	p_0=0.901 (p_1=0.099), ω_0=0.077 (ω_1=1)
M2a (positive selection)	-2190.27	0.213	p_0=0.936, p_1=0.056 (p_2=0.009), ω_0=0.105 (ω_1=1), ω_2=6.802
M7 (beta)	-2197.92	0.184	p=0.098, q=0.437
M8 (beta&ω)	-2190.51	0.212	p_0=0.991 (p_1=0.009), p=0.431, q=2.312, ω_s=6.740
B/YM-lineage (II-*iv*) 1994~2005 (33 strains)			
M0 (one-ratio)	-2144.23	0.220	ω=0.220
M1a (nearly neutral)	-2127.98	0.211	p_0=0.789 (p_1=0.211), ω_0=0 (ω_1=1)
M2a (positive selection)	-2127.40	0.242	p_0=0.847, p_1=0.022 (p_2=0.131), ω_0=0.021 (ω_1=1), ω_2=1.537
M7 (beta)	-2128.03	0.200	p=0.005, q=0.020
M8 (beta&ω)	-2127.40	0.242	p_0=0.868 (p_1=0.132), p=0.052, q=1.023, ω_s=1.560
B/VI-lineage (III-*i*) 1975~1993 (24 strains)			
M0 (one-ratio)	-2530.27	0.336	ω=0.336
M1a (nearly neutral)	-2492.37	0.251	p_0=0.749 (p_1=0.251), ω_0=0 (ω_1=1)
M2a (positive selection)	-2477.70	0.351	p_0=0.736, p_1=0.254 (p_2=0.010), ω_0=0.005 (ω_1=1), ω_2=9.700
M7 (beta)	-2492.81	0.222	p=0.009, q=0.029
M8 (beta&ω)	-2477.89	0.354	p_0=0.983 (p_1=0.017), p=0.016, q=0.050, ω_s=7.267
B/VI-lineage (III-*ii*) 1996~2007 (98 strains)			
M0 (one-ratio)	-2924.64	0.299	ω=0.299
M1a (nearly neutral)	-2899.56	0.266	p_0=0.805 (p_1=0.195), ω_0=0.088 (ω_1=1)
M2a (positive selection)	-2887.74	0.320	p_0=0.796, p_1=0.199 (p_2=0.005), ω_0=0.094 (ω_1=1), ω_2=9.871
M7 (beta)	-2899.97	0.266	p=0.156, q=0.430
M8 (beta&ω)	-2887.00	0.309	p_0=0.994 (p_1=0.006), p=0.244, q=0.698, ω_s=8.534

Early strain (I) (1940~1970). Among the 271 HA_1 sequences analyzed in this study, a total of 11 sequences over a time span of 31 years belong to this group (**Fig.2.2, Fig.S1a**). To limit the uncertainties related to the relatively small number of samples in this group, the results from Bayes Empirical Bayes analysis were used throughout this study [35, 42]. In LRT tests, the values of $2\Delta l$ were 5.36 for M2a versus M1a, and 5.76 for M8 versus M7 (**Table 2.2**). These values were larger than the critical value of $\chi^2_{5\%} = 3.84$, but smaller than $\chi^2_{1\%} = 6.63$ with d.f. = 1 [7, 25, 32, 35]. The M2a model suggested ~0.5% sites to be under positive selection with ω_2=7.990 (**Table 2.1**). Similarly, the M8 model suggested ~0.6% sites to be under positive selection with ω_s=7.428. The M2a model identified a total of six sites to be under positive selective pressure (>50% posterior probability) (**Table 2.3**). The M8 model identified 14 sites of being under positive selective pressure (>50% posterior probability) (**Fig.2.3a**). Among them, two sites were of greater than 95% posterior probability to be under positive selection: HA_1 167 (95%) on the 160-loop and 194 (99%) on the 190-helix.

B/YM-like lineage (II). A total of 138 HA_1 sequences in this analysis belong to B/YM-like lineage. It was further divided into four sublineages, II-*i* (25 sequences), II-*ii* (24 sequences), II-*iii* (56 sequences) and II-*iv* (33 sequences) (**Fig.2.2**).

Table 2.2. Likelihood ratio tests (LRT) between M2a versus M1a and M8 versus M7 for the seven subgroups of HA_1 subunit of influenza B virus strains circulating between 1940~2007

LRT	$2\Delta\ell$ *
Early strain (I) 1940~1970 (11 strains)	
M2a – M1a	5.36
M8 – M7	5.76
B/YM-lineage (II-*i*) 1972~1984 (25 strains)	
M2a – M1a	44.44
M8 – M7	46.14
B/YM-lineage (II-*ii*) 1987~1996 (24 strains)	
M2a – M1a	14.50
M8 – M7	14.26
B/YM-lineage (II-*iii*) 1991~2002 (56 strains)	
M2a – M1a	12.22
M8 – M7	14.82
B/YM-lineage (II-*iv*) 1994~2005 (33 strains)	
M2a – M1a	1.16
M8 – M7	1.26
B/VI-lineage (III-*i*) 1975~1993 (24 strains)	
M2a – M1a	29.34
M8 – M7	29.84
B/VI-lineage (III-*ii*) 1996~2007 (98 strains)	
M2a – M1a	23.64
M8 – M7	25.94

* In LRT tests, the values of $2\Delta\ell$ were compared with the critical values of χ^2 distribution (6.63 and 3.84 for $\chi^2_{1\%}$ and $\chi^2_{5\%}$, respectively, with d.f.=1) [7, 25, 32, 35]. Significantly larger values of $2\Delta\ell$ over those of χ^2 distributions led to the rejection of the null models M1a and M7.

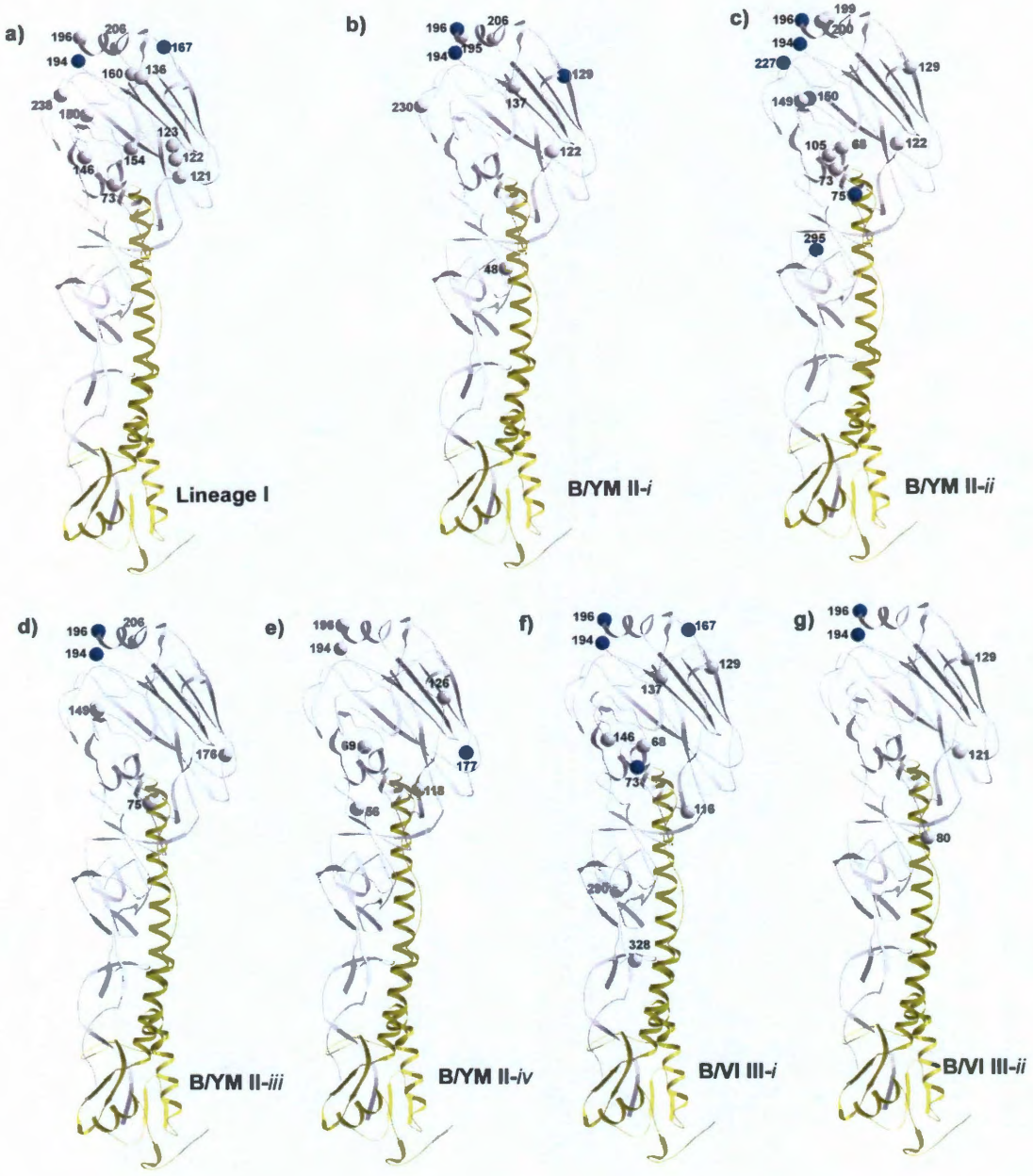


Fig.2.3. Sites with posterior probabilities of greater than 50% to be under positive selection in the M8 models for the seven subgroups of influenza B virus HA, in the order of early strain (I) (a), B/YM-lineage (II-*i* ~ II-*iv*) (b~e) and B/VI-lineage (III-*i* and III-*ii*) (f, g). Each site is shown as a ball centered at its C□ atom in the structure (Protein Data Bank code 3BT6) [30]. Sites with greater than 95% posterior probability to be under positive selection are shown in dark color and the rest are in light color. The structure of one monomer of HA is in the same orientation as the monomer shown in color in Fig.2.1.

Early strain sublineage (II-*i*) (1972~1984). These early strains of B/YM-lineage spanned a period of 13 years (**Fig.S1b**). The $2\Delta l$ values of M2a versus M1a and M8 versus M7 were much greater than $\chi^2_{1\%} = 6.63$ with d.f. = 1 (**Tables 2.1, 2.2**), resulting in the rejection of the null models M1a and M7. Both M2a and M8 models suggested ~1.1% sites to be under strong positive selection with large ω values (**Table 2.1**). The M2a model identified a total of five sites to be under positive selection (>50% posterior probability) (**Table 2.3**), three of which were of greater than 95% posterior probability: HA_1 129 (97%) on the 120-loop, 194 (100%) and 196 (100%) on the 190-helix. These three sites were again with >95% posterior probability in the M8 model: HA_1 129 (99%), 194 (100%) and 196 (100%) (**Fig.2.3b**).

Sublineage (II-*ii*) (1987~1996). The B/YM-lineage strains in this group covered a 10-year period (**Fig.S1c**). In LRT tests, the $2\Delta l$ values of M2a versus M1a and M8 versus M7 provided strong support for the existence of positive selection (**Tables 2.1, 2.2**). Both M2a and M8 models suggested ~4% sites to be under positive selection with $\omega \approx 4$ (**Table 2.1**). The M2a model identified two sites with higher than 95% posterior probability of being positively selected (**Table 2.3**): HA_1 194 (99%) and 196 (97%) on the 190-helix. The M8 model identified a total of six sites with greater than 95% posterior probability of being positively selected (**Table 2.3**): HA_1 75 (97%) and 295 (98%) on the 120-loop, 150 (96%) on the 150-loop, 194 (100%), 196 (99%) and 227 (97%) on the 190-helix (**Fig.2.3c**). It is noteworthy that HA_1 150 on the 150-loop was inferred to be under positive selection

with very high confidence, in excellent agreement with previous conclusions that the 150-loop is an important epitope for B/YM-lineage [26, 29].

Sublineage (II-*iii*) (1991~2002). This sublineage of B/YM-like strains covered a 12-year period (**Fig.S1d**). The LRT tests led to the rejection of the null models M1a and M7 (**Tables 2.1, 2.2**). Both M2a and M8 models suggested ~0.9% sites to be under positive selection with $\omega \approx 7$ (**Tables 2.1, 2.2**). The M2a model revealed three sites of being under positive selection, including HA_1 194 (97%) and 196 (99%) on the 190-helix (**Table 2.3**). These two sites were of 99% and 100% posterior probability of positive selection in the M8 model (**Table 2.3 and Fig.2.3d**).

Sublineage (II-*iv*) (1994~2005). This sublineage of B/YM-like strains contained some of the most recently circulating strains of B/YM-lineage (**Fig.S1e**). In sharp contrast to all other sublineages of B/YM-like strains and to all B/VI-like strains, the LRT statistics were $2\Delta l = 1.16$ and 1.26 for M2a versus M1a and M8 versus M7, respectively (**Table 2.3**), suggesting a low confidence for the existence of positive selection. In Bayes Empirical Bayes analysis, both M2a and M8 models suggested a relatively large percentage of sites (~13%) to be under very weak positive selection with ω_2=1.537 and ω_s=1.560, respectively (**Table 2.1**). The M2a model identified a total of four positively selected sites with >50% posterior probability (**Table 2.3**). In the M8 model, a total of seven sites were identified, with only one site, HA_1 177 (98%) on the 120-loop, with > 95% posterior probability (**Table 2.3 and Fig.2.3e**). This sublineage was the only group in which HA_1 194 and 196 are of lower than 95% probability to be under positive selection.

B/VI-like lineage (III). A total of 122 HA_1 sequences of influenza B virus strains belong to this lineage. They were grouped into two sublineages, early strains (III-*i*) containing 24 sequences and more recent strains (III-*ii*) containing 98 sequences (**Fig.2.2**).

Early strain sublineage (III-*i*) (1975~1993). These early strains of B/VI-lineage spanned a time period of 19 years and exhibited significant sequence differences from the recent circulating B/VI-like strains (III-*ii*) (**Fig.2.2, Fig.S1f**). The LRT statistics supported the rejection of the null models (M1a and M7) and strongly supporting the presence of positive selection (**Tables 2.1, 2.2**). The M2a model suggested 1.0% sites to be under positive selection with ω_2=9.700 (**Table 2.1**). Similarly, the M8 model suggested 1.7% sites to be under positive selection with ω_s=7.267. The M2a model identified seven sites to be under positive selective pressure (**Table 2.3**), with HA_1 194 (100%) and 196 (100%) on the 190-helix of higher than 95% posterior probability. The M8 model revealed a total of 11 positively selected sites, among which four sites were of greater than 95% posterior probability. They were HA_1 73 (95%) (120-loop), 167 (97%) (160-loop), 194 (100%) and 196 (100%) (190-helix) (**Table 2.3, Fig.2.3f**). It is important to note that among these four sites with the highest posterior probability, one site, HA_1 167, is on the 160-loop, while none is located on the 150-loop. In sharp contrast, B/YM-like (II-*ii*) sublineage, which circulated in an overlapping time period and contained the same number of sequences, had HA_1 150 on the 150-loop to be under positive selection. These observations further supported earlier conclusions that the 160-loop epitope is

specific for the B/VI-lineage strains [27-28] while the 150-loop epitope is specific for the B/YM-lineage strains [26, 29].

Recent strain sublineage (III-*ii*) (1996~2007). The more recent isolates of B/VI-lineage strains remained to be a single group over the time period of 12 years (**Fig.S1g**). The LRT statistics supported strongly the presence of positive selection (**Tables 2.1, 2.2**). The M2a model suggested ~0.5% sites to be under positive selection with ω_2=9.871 (**Table 2.1**). Similarly, the M8 model suggested ~0.6% sites to be under positive selection with ω_s=8.534. HA_1 194 had a posterior probability of 95% and 99% to be under positive selection in M2a and M8 models, respectively, while HA_1 196 has a 100% posterior probability in both M2a and M8 models (**Table 2.3 and Fig.2.3g**). Compared to the earlier B/VI-like (III-*i*) sublineage, one noticeable difference is that the 160-loop was no longer under positive selection in these recent strains (III-*ii*). Rather, positive selection was focused on the 190-helix.

Table 2.3. Sites with higher than 50% posterior probabilities of being under positive selective pressure for the HA_1 subunit of influenza B virus strains circulating between 1940~2007

Model	Positively selected sites[1]
	Early strain (I) 1940~1970 (11 strains)
M2a (positive selection)	73, 150, 167*, 194**, 196, 238
M8 (beta&ω)	73*, 121, 122, 123, 136, 146, 150*, 154*, 160, 167***, 194***, 196**, 206, 238*
	B/YM-lineage (II-*i*) 1972~1984 (25 strains)
M2a (positive selection)	122, 129***, 194***, 196***, 206
M8 (beta&ω)	48, 122**, 129***, 137, 194***, 195, 196***, 206**, 230
	B/YM-lineage (II-*ii*) 1987~1996 (24 strains)
M2a (positive selection)	68*, 75**, 122, 129, 150**, 194***, 196***, 200, 227**, 295**
M8 (beta&ω)	68**, 73, 75***, 105, 122*, 129**, 149, 150***, 194***, 196***, 199, 200, 227***, 295****
	B/YM-lineage (II-*iii*) 1991~2002 (56 strains)
M2a (positive selection)	176, 194***, 196***
M8 (beta&ω)	75, 149, 176*, 194***, 196***, 206
	B/YM-lineage (II-*iv*) 1994~2005 (33 strains)
M2a (positive selection)	69*, 177**, 194, 196
M8 (beta&ω)	56, 69**, 118, 126, 177***, 194**, 196**
	B/VI-lineage (III-*i*) 1975~1993 (24 strains)
M2a (positive selection)	73*, 116, 167**, 194***, 196***, 290, 328
M8 (beta&ω)	68, 73***, 116**, 129, 137, 146, 167***, 194***, 196***, 290**, 328**
	B/VI-lineage (III-*ii*) 1996~2007 (98 strains)
M2a (positive selection)	194***, 196***
M8 (beta&ω)	80, 121*, 129, 194***, 196***

[1]Positively selected sites from Bayes Empirical Bayes analysis [35].
*Posterior probability of positive selective pressure is between 75~84%.
**Posterior probability of positive selective pressure is between 85~94%.
***Posterior probability of positive selective pressure is higher than 95%.

2.4 Discussion

Roles of antibody selection in the evolution of influenza B virus HA

In previous studies, four major antigenic epitopes of influenza B virus HA, the 120-loop, the 150-loop, the 160-loop, and the 190-helix, were identified on the membrane-distal domain of HA_1 [30] (**Fig.2.1**). Strikingly, in this study, all the

identified positively selected sites in the seven subgroups were located on these four major antigenic epitopes, supporting the important roles of antibody selection in the molecular evolution of influenza B virus HA.

The 150-loop is an important epitope on HA. Antigenic properties were altered for influenza B virus with mutations on this loop in laboratory-selected escape mutants [15-16, 18, 26], field isolates [29, 43] and egg-adapted variants [44-47]. In more recent influenza B virus isolates, the 150-loop region appeared to be the neutralizing epitope specific for B/YM-like strains [26, 29]. Consistent with that finding, HA_1 150 was under positive selection with 96% posterior probability in B/YM-like (II-*ii*) sublineage.

The 160-loop is the only region in influenza B virus HA where insertions, deletions and single amino-acid substitutions were detected in field isolates [28, 37, 48] and mAb-escape mutants [15-16, 18, 27, 29], as an effective way for influenza B virus to survive a long period of time without antigenic shifts as observed in influenza A virus [37]. In recent isolates, the 160-loop became specific for B/VI-like lineage [27-28]. In agreement with this observation, HA_1 167 on the 160-loop was selected positively in early strains (I) and in B/VI-like (III-*i*) sublineage (**Table 2.3, Fig.2.3**), with 95% and 96% posterior probability, respectively.

The 190-helix, which forms part of the receptor-binding site (RBS) of influenza B virus HA, is inarguably one of the most important epitopes. The hot spot is at HA_1 194~196, a potential glycosylation site. Similar to influenza A virus HA [49-53], influenza B virus HA also utilized the addition or removal of glycosylation as a mechanism for antigenic drift [15-16, 30, 45-46, 54-62]. In this current analysis,

HA_1 194 and 196 were constantly identified to be under positive selective pressure, with greater than 99% probability in 11 out of 14 cumulative cases (combining both sites in seven groups), and over 85% in three other cases. HA_1 227 in sublineage (II-*ii*) was another positively selected site on 190-helix with high posterior probability (97%) (**Table 2.3, Fig.2.3**).

Perhaps one of the most important observations from this study is positive selection of the 120-loop region. The 120-loop epitope was defined as HA_1 116~137 and its surrounding regions [30]. In this context of this article, we refer to all sites not adjoining the 150-loop, 160-loop or the 190-helix epitopes as the 120-loop region due to spatial proximity (**Fig.2.1**). Although the 120-loop region appeared to be one of the most frequently mutated regions in field isolates [14], its role in antigenicity of influenza B virus HA was not recognized until most recently [30, 47, 63]. One possibility for such a delay in recognition is that the 120-loop is proximal to the viral envelope membrane, making the access by antibodies more difficult, as observed for influenza A virus HA [64-69]. Thus, it is very important that this current study provided strong evidence for positive selection of the 120-loop region, further supporting its significance in antigenicity of influenza B virus HA.

Trends of positive selection on influenza B virus HA

The early strains (I) seemed to have rather even distribution of positive selective pressure on all four major epitopes, although the positive selection on the 160-loop and 190-helix appeared to be stronger and/or more prevailing (**Table 2.3, Fig.2.3a**). In contrast, the early strains of B/YM-lineage and B/VI-lineage,

sublineages (II-*ii*) and (III-*i*) respectively, were sharply divided. HA_1 150 on the 150-loop in B/YM-like (II-*ii*) sublineage, and HA_1 167 on the 160-loop in B/VI-like (III-*i*) sublineage, were inferred to be under positive selection with high posterior probability (**Table 2.3**). These observations agreed very well with earlier studies in which the 150-loop and 160-loop were found to be specific epitopes for the B/YM- and B/VI-lineages, respectively [26-29]. However, despite large sequence differences, the recent B/YM-like (II-*iii*) sublineage and B/VI-like (III-*ii*) sublineage converged at focusing on the 190-helix for antigenic drift (**Table 2.3 and Fig.2.3**). Most strikingly, in the newest B/YM-like (II-*iv*) sublineage, a large number of sites were found to be under rather weak positive selection, and the only positively selected site identified with high confidence was HA_1 177 on the 120-loop. The new trends of positive selection among these most recent strains, in conjunction with results from other studies [30, 47, 63], stress the increasingly important role of the 120-loop in antigenicity of influenza B virus HA.

Concluding remarks

This study reports a large-scale systematic analysis of diversifying positive selective pressure on HA of distinct lineages/sublineages of influenza B virus isolated in the past 68 years. The highlights of the results from this study are: **a).** The number of positively selected sites in influenza B virus HA were much fewer than those of influenza A virus HA [23]; **b).** Although it does not have subtypes as influenza A virus HA, influenza B virus HA did and continue to diverge into different sublineages. This was particularly true for B/YM-lineage, as exemplified by the

newly emerging B/YM-like (II-*iv*) sublineage that had not been previously described. **c).** The study revealed the predominant roles of antibody selection in the molecular evolution of influenza B virus HA. **d).** Despite the differences among different lineages/sublineages, HA_1 194 and 196 were constantly under positive selective pressure in all but one cases. **e).** The 120-loop was an important epitope under constant positive selection. It may play an increasingly important role in antigenicity in future field isolates, as evidenced in the most recent B/YM-like (II-*iv*) sublineage (**Table 2.3**). **f).** Each lineage/sublineage utilized their respective favorite sites in positive selection. Thus, for any newly emerging strains of influenza B virus, it is important to put them in the context of their evolutionary history in order to understand and appreciate their full epidemic potential.

Chapter 3. Polycomb Repressive Complex 2

3.1 Background

There exist two types of epigenetic modifications: DNA methylation and post-translational modifications on histone tails in the form of methylation (one, two, or three), acetylation, and ubiquitination on lysines or arginines, or phosphorylation of serines, leading to repressive or activated gene expression states [70-71]. However, the principles and mechanisms underlying epigenetic regulation remains one of the largest mysteries in our understanding of cellular programming. This is particularly true regarding how these processes are targeted toward specific genes in a temporal/spatial dependent manner, allowing for the complex process of cellular differentiation, or in the aberrant case, oncogenic transformation, to occur. Among them, one process under active investigation is the targeting and trimethylation of histone 3 lysine 27 (H3K27me3) by Polycomb Repressive Complex 2 (PRC2) to promote a repressive state in neighboring genes[71].

Originally discovered in *Drosophila* as the regulators of homeotic (HOX) genes, polycomb group (PcG) proteins are well-conserved epigenetic modifiers that repress the expression of thousands of target genes in a given genome [72-83]. These target genes are essential for many fundamental, evolutionarily conserved processes including development, cell fate determination, proliferation, stem cell pluripotency and self-renewal [72, 75, 78-79, 84-87]. Mutations of PcG proteins are implicated in defects in stem cell fates and their abnormal levels exhibit a striking correlation with the severity and invasiveness of a number of cancer types including prostate cancer

and breast cancer [72, 75, 78-79, 84-87].

PcG proteins impose gene silencing through their interactions with polycomb response elements (PREs) that are present on the promoter regions of polycomb target genes [88]. This interaction is mediated by three types of multiprotein complexes, polycomb repressive complex 1 and 2 (PRC1 and PRC2) and a recently discovered PhoRC that contains the DNA-binding protein Pleiohomeotic (Pho) or Pleiohomeotic-like (PhoL) [89] in *Drosophila* and Ying and Yang 1 and 2 (YY1 and YY2) in mammals [90-92]. The known members of *Drosophila* PRC1 include Polycomb (PC), Polyhomeotic (PH), Posterior sex combs (PSC) and dRing, while *Drosophila* PRC2 contains at least three core components: Enhancer of zeste (E(z)), Extra sex comb (Esc), and Suppressor of zeste 12 (Su(z)12) [90]. Since none of these PRC1 and PRC2 proteins can bind to DNA directly, a hierarchical recruitment model has been proposed stating that DNA-binding transcription factors including Pho and PhoL first bind to PREs on the target genes and recruit the PRC2 complex to trimethylate the lysine 27 residue of histone H3 (H3K27me3) that is later bound by the PRC1 complex for maintenance [93]. Besides Pho and PhoL, the best studied *Drosophila* transcription factors contributing to PRC2 recruitment include GAGA factor (GAF)/Pipsqueak (PSQ) [94-96], Zeste [97-98], Dorsal switch protein (DSP) [99-100], Grainyhead (Grh) [101] and Sp1/KLF [102] (reviewed by Ringrose and Paro [103]). In addition, several *Drosophila* PREs have been identified through both computational and experimental analyses [104-113]. More recently, the first two mammalian genomic regions have been discovered to confer PcG responsiveness, one

in the human HOXD cluster [114] and the other in the regulatory region of the mouse MafB gene [80].

Recent advances in high-throughput techniques such as chromatin immunoprecipitation in conjunction with microarray (ChIP-on-chip), DNA adenine methyltransferase identification (DamID) and ChIP-sequencing (ChIP-seq), have greatly enriched our knowledge on the scale of genes regulated by PcG [72, 75-79, 81-82, 84-87, 115-122]. However, the rather low overlaps of target genes identified in separate ChIP studies, at approximately 30% for three ChIP studies on *D. melanogaster* [75, 85-86, 103], stress the need for additional experimental and computational verifications of individual PcG target genes.

This is perhaps even a more difficult problem in human biology. In order to shed light on mammalian PREs and their target genes, more than 22 different whole genome experiments across 16 different cell lines have been reported, utilizing CHIP-CHIP or CHIP-seq techniques to map the PRC2 component proteins (Suz12, EED, EZH2) or the H3K27me3 mark to specific genomic loci or to the promoter regions of known human genes (~24,000) [72, 76-79, 81-82, 87, 116-123] These experiments have yielded hundreds to thousands of genes that could be potential PRC2 targets. However, the identification of new *bona fide* PREs based on these individual CHIP experiments has proven difficult as CHIP experiments represent an ensemble average or all of the cells in the experiment and as such are subject to potential biases. They also fail to account for the initiation of PRC2 binding, and instead, focus mostly of established binding domains, which could be subject to spreading from other initiation loci. Finally, analysing for mammalian PRE's are made even more difficult

by the lack of established DNA binding proteins meaning that it is currently only possible to apply top-down methods of modelling, whereas in Drosophila it has been demonstrated that bottom up approaches are least possible, even if they are not as accurate as one would like. Ideally, a powerful computational method that is able to predict/screen, with a reasonable accuracy, PcG target genes in a given genome would drastically expedite experimental verification of these genes.

In the literature, there are considerable efforts in developing computational methods to predict PRE sequences and to locate the genes regulated by PcG based upon their adjacency to PREs. For instance, Ringrose *et al.* investigated the combinatorial pattern of transcription factors known to be involved in PcG recruitment and assigned to each genomic region of interest a score equaling the weighted sum of the occurrence of every possible transcription factor pairs [113]. Fiedler and Rehmsmeier extended this idea and developed *jPREdictor* for PRE prediction [124]. Hauenschild and colleagues used the latest version of *jPREdictor* to perform a genome-wide prediction on *D. melanogaster* and predicted 201 PREs together with 243 associated genes [125]. They also incorporated the aspect of comparative genomics and expanded their prediction to 285 PREs with 322 associated genes. More recently, Liu *et al.* integrated data from a ChIP study and transcription factor binding analysis to predict a set of PcG target genes in mouse embryonic stem cells [126]. Despite these efforts, however, due to the plasticity of PRE sequences, developing a reliable computational PRE predictor remains a difficult task. For example, the overlaps between the top target genes predicted by

jPREdictor and those shown in the three recent ChIP studies in *D. melanogaster* [75, 85-86] are strikingly low (at 8%~20%).

We have addressed this challenge by developing a novel computational approach, *EpiPredictor*, to predict PcG target genes via the identification of PRE sites. With the incorporation of novel features including the use of a support vector machine (SVM)-based classifier, global sequence information, conservation analysis and comparative genomics, our approach was able to predict PcG target genes in the *D. melanogaster* genome with substantially improved accuracy. Most of the predicted PcG target genes are transcription factors involved in key biological processes such as development, neurogenesis and cell fate determination. Our results suggest that multiple transcription factor networking at the *cis*-regulatory elements is critical for PcG recruitment, and high GC content and high conservation level are also important features of PcG target genes.

However, in order to be able to apply these same techniques into human cells we must first try to identify the motifs and genomic features that are the most applicable to human cell biology. With the wealth of new genome-wide information on potential PRC2 target genes, we reasoned that, by searching for conserved motifs amongst the genes with high CHIP signals across many different experimental conditions, the amount of noise can hopefully be reduced to a level at which meaningful results might be obtained. Since PREs in general contain multiple conserved motifs that attract transcription factors [113], the conserved DNA motifs may in turn be used as input for computational search of potential PREs that can be further experimentally verified. To this end, we applied multiple advanced

computational methods to published CHIP data and obtained a list of conserved DNA motifs and compared them against the known transcription factor binding motifs found in the Jaspar (human) [127], and select motifs from the Transfac (human) [128] database and motif localization data from Oncomine[129]. Moreover, we have identified two distinct classes of motifs that are differentially enriched in cancer cells and in embryonic stem cells. The implications of these findings are also discussed.

3.2 Materials and Methods

3.2.1 Methods for Drosophila Modeling

Selection of motifs

In *Drosophila*, several transcription factors responsible for PcG recruitment have been identified, which, together with the consensus sequences of their DNA binding sites, are collectively referred to as *motifs* hereafter. We used seven motifs corresponding to four transcription factors, GAF (G, G10), Pho (PS, PM, PF), engrailed (EN1), and Zeste (Z), all of which are known to be instrumental for PcG recruitment (3.1). The same motif set was also used in *jPREdictor* [113, 125]. Though a few other transcription factors, *e.g.*, DSP, Grh, Sp1/KLF, are also implicated in PcG recruitment in some studies, we did not include them in our current system because doing so did not lead to any performance improvement (data not shown) and also may not allow a fair comparison with *jPREdictor*.

Table 3.1. Motifs for transcription factors used for prediction model.

Alias	Motif	Max Mutation Allowed
G	GAGAG	0
G10	GAGAGAGAGA	1
PS	GCCAT	0
PM	NCGCCATNDNND	0
PF	GCCATHWY	0
EN1	GSNMACGCCCC	1
Z	YGAGYG	0

Construction of the validation sets

To validate our prediction of PcG target genes in an objective way, we used the gene lists reported in three recent ChIP studies in *D. melanogaster*, where Schwartz *et al.* [75] used ChIP-on-chip technique on *S2* cultured cell line with antibodies to PC, E(z), PSC and H3K27me3; Tolhuis *et al.* [86] used DamID approach on *Kc* cells to identify binding sites of PC, Esc, Sex combs extra (Sce) and H3K27me3; while Schuettengruber *et al.* [85] applied ChIP-on-chip on *Drosophila* embryos and employed antibodies to PC, PH and H3K27me3. Different choices of cell lines and antibodies all had an impact on the results of these experiments that differed from one another at varying degrees. Since our *in silico* PRE prediction is independent of any experimental conditions, we expected that a comparison of our results with these three well-annotated studies, which as a whole investigated a range of antibodies and cell types, would provide a comprehensive evaluation of our system. To ensure that the validation gene lists to be used were as reliable and up-to-date as possible, we performed a post-processing procedure on the published data using the following stringent selection criteria. For all three validation sets, we used

the gene lists published by the authors as input and removed duplicates if there was any. We also eliminated the genes that were withdrawn in the newer release of the gene annotation to ensure that the validation gene sets are up-to-date. In particular, in processing Schwartz's data [75], we only selected the target genes with strong PcG binding signals to all of the four PcG proteins (PC, PSC, PH and Su(z)2) simultaneously as defined by the authors. As a result, we obtained three lists consisting of 176 (Schwartz), 225 (Tolhuis) and 215 (Schuettengruber) predicted PcG target genes, respectively (**Appendix Table 4.1**). Among them, 38 genes appeared in all of the three validation sets, denoted as *Intersection*, making the degree of overlap in the range of 17%~22%.

Construction of the training set

Our PRE classifier is a supervised learner. Therefore we needed to provide it with a training set of good quality. This consisted of two steps: 1) construction of a PRE/non-PRE sequence collection and 2) construction of the training set containing examples of both PRE sites (positive) and non-PRE sites (negative).

First we constructed a sequence collection containing 12 known PRE sequences and 23 control (non-PRE) sequences. Among them, the 12 PRE sequences and 16 control sequences were the same as those used by Ringrose and colleagues [113, 125]. The 12 PRE sequences had solid evidence to support the existence of PRE site(s) within while the 16 control sequences included promoters of genes regulated by GAF and Zeste but not by PcG proteins [113]. To reflect the most recent progress in the field, we followed the same methodology used by Ringrose

[113] and collected seven extra control sequences (2L:131425..131940; 2L:16715553..16716125; 3R:2949268..2950339; 3R:5338373..5339021; 3R:12279006..12279593; 3R:12879696..12880258; and 3R:19931380..19931932) for our training set that were bound by GAF, Pho and Zeste but did not have any enrichment for PC, PH or H3K27me3 in a genome-wide ChIP study [85]. They were obtained by examining whether a given locus bound by GAF, Pho and Zeste was in the proximal promoter region of any gene, *i.e.*, -1,000 to +1,000 base pairs (bps) with respect to the gene's transcription start site (TSS). If so, we retained the locus and the gene, otherwise, we discarded them. To ensure that our control sequences did interact with GAF, we consulted another list of GAF target genes by an independent study [130]. If the genes associated with any retained loci under investigation were not included in the second study, the loci were eliminated from our list. It was evident that, despite the addition of seven new control sequences in our study, the size of the sequence collection remained rather small.

A PRE sequence containing PRE site(s) is much larger than an actual PRE site. Due to the limited resolution of the experimental verification process, most known PRE sequences included in our sequence collection spanned thousands of bps long whereas the core-PRE sites are usually much shorter (<200 bps) [110]. In other words, in addition to core PRE sites, a known PRE sequence might also contain non-PRE sites. Thus it was prudent to identify the loci that were most likely the *bona fide* PRE sites. For this purpose, we scanned each PRE/non-PRE sequence in our collection with a sliding window of 200 bps that incrementally moved downstream with a constant step of 20 bps. For each PRE sequence, we chose the window(s) with

the highest sum of motif occurrence (calculated by the *Motif Analyzer* in the following section) as PRE sites. For every control (non-PRE) sequence, all the windows from scanning a control sequence were kept to ensure that the classifier was to be trained under very stringent condition.

Our new system *EpiPredictor*

Our system consisted of six primary components including: *Motif Analyzer*, *PRE Classifier*, *GC Analyzer*, *PRE-to-gene Mapper*, *Conservation Level Analyzer* and *Comparative Genomics Analyzer* (**Fig.3.1a**). With the exception of *PRE-to-gene Mapper*, which was a utility module, each component rendered a unique perspective of investigating the genomic sequence or gene of interest. The first three units were centered around the prediction of PRE sites (**Fig.3.1a, b**) whereas the last three were focused on analyses at the gene level (**Fig.3.1a**).

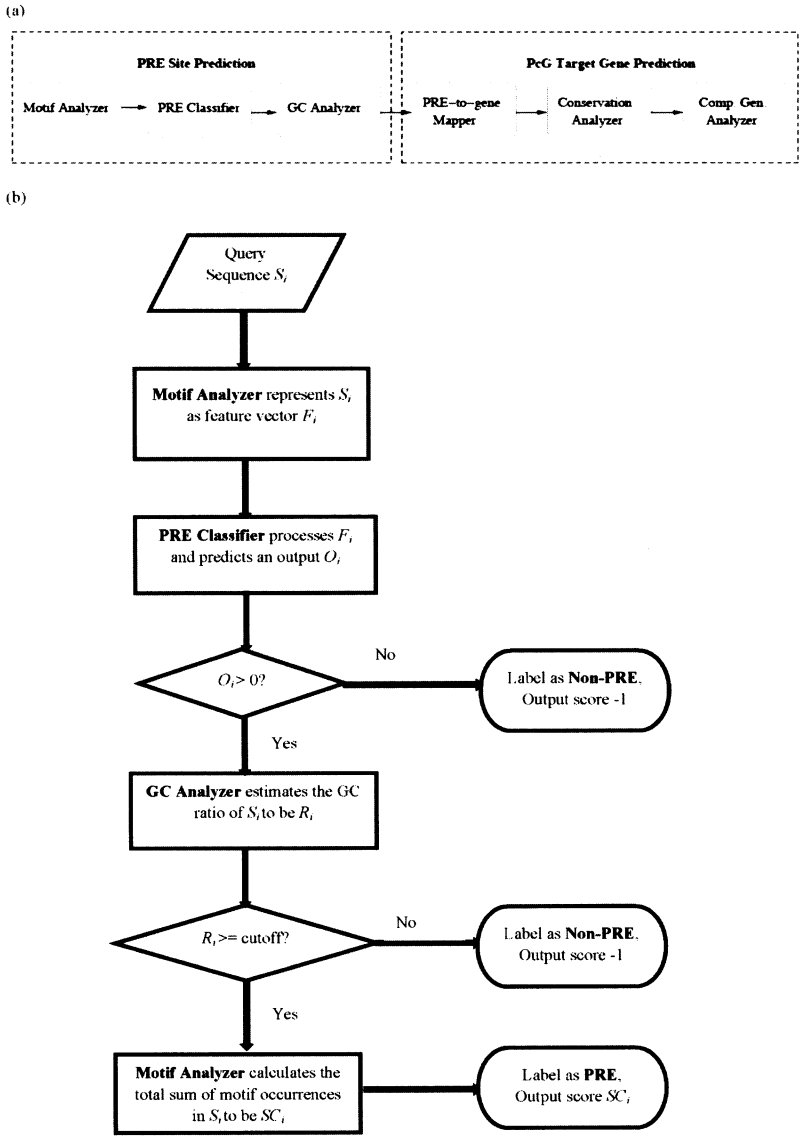


Fig.3.1. The *EpiPredictor* system. A) Architecture of *EpiPredictor*. The modules of Motif Analyzer, PRE Classifier and GC Analyzer are dedicated to the prediction of PRE sites and those of PRE-to-gene Mapper, Conservation Level Analyzer and Comparative Genomics Analyzer are focused on the prediction of PcG target genes. B) Flowchart of the PRE site prediction modules of *EpiPredictor*.

Prediction of PRE sites

Motif Analyzer

We employed a sliding window of 200 bps and a step size of 20 bps to scan the whole genome where the DNA sequence overlapping with the window was captured and analyzed at any given time by the Motif Analyzer. Using a set of n motifs of transcription factors that were known to be involved in PcG recruitment, denoted by $M_1, M_2, \ldots, M_n$, the Motif Analyzer constructed a profile for each window sequence/locus (denoted by S_i) and represented it by a feature vector $F_i = (f_{i1}, f_{i2}, \ldots, f_{in})$, where f_{ij} denoted the occurrence frequency of motif M_j in sequence S_i. This feature vector was then analyzed by the pre-trained PRE classifier (below) that predicted whether the test window/locus was a PRE or not (**Fig.3.1b**).

SVM-based PRE Classifier

Ringrose *et al.* [113] examined the occurrence of paired motifs at the putative PRE sites and observed that the weighted sum of the occurrence frequencies of all possible motif pairs were far more effective than a linear sum up of the occurrence frequency of single motifs. This suggested that the pattern of transcription factor interactions at the PRE sites be combinatorial. In order to abstractly model the multifaceted interactions among transcription factors at PRE sites, we incorporated an SVM-based PRE classifier, which is a powerful supervised learning method for handling classification tasks. SVM has achieved prominent success in a spectrum of biological applications including gene selection [131-132], protein classification [133-135], cancer tissue characterization [136-137], outperforming many other classic

machine learning techniques such as neural network, decision tree, *k*-nearest neighbor [138-139].

There are four basic kernel functions in SVM, including linear, polynomial, radial basis function (RBF) and sigmoid. Given the context of PRE prediction, we provided a further annotation to SVM coupled with some of these kernels. For instance, in the case of a polynomial kernel, the parameter d corresponded to the degree of motif combinations, *e.g.* when $d = 1$ (equivalent to a linear kernel), only single motif occurrence was noted; when $d = 2$ (quadratic kernel), the occurrence of motif pairs was considered; whereas when $d = 3$ (cubic kernel), the occurrence of motif triplets was analyzed. In the case of the RBF kernel, the data was mapped to an infinite dimensional Hilbert space where intuitively speaking, all the motifs were mapped to a circle/hypersphere. Taken together, we expected the polynomial ($d > 1$) and RBF kernels to be best for modeling transcription factor interaction at the PRE sites. While the windows/loci classified to be non-PREs were discarded, those classified to be PREs had to undergo further scrutiny by GC Analyzer (below).

GC Analyzer

Previous studies indicated that native DNA sequence features, such as GC content, are associated with epigenetic modification activities such as DNA methylation and PcG binding [140-142]. In particular, the work of Ku *et al.* [76] suggested that CpG islands influence recruitment of PcG. Furthermore, GC-rich sequence elements have been shown to recruit PRC2 in mammalian embryonic stem cells [143] and high-CpG-density promoters are associated with highly regulated key developmental genes and are enriched with the H3K27me3 marks [144]. Therefore,

we implemented a GC Analyzer to further scrutinize the output from the PRE Classifier. For a sequence window/locus S_i that was positively predicted by the PRE classifier, our GC Analyzer compared its GC ratio R_i with a threshold value R_T and discarded S_i if $R_i < R_T$ (**Fig. 3.1b**). To decide on an appropriate threshold for a region of 200 bps window size that we used, we examined six experimentally verified PRE sequences where short core PRE segments were identified [110]. The lowest GC ratio of these core PRE segments was 44%. We then chose this lower bound of the GC ratio as our threshold in order to ensure that all the verified PRE segments satisfy this GC ratio cut-off so that they can pass the GC Analyzer's scrutiny. We also compared the cut-off values of 44%, 42% or 40% on *EpiPredictor*, and found 44% yielded the best performance. Therefore, we used the 44% threshold in our subsequent analysis. Only the ones that passed the GC content test were considered as the potential PRE loci. Each locus was given a numerical score SC_i by the Motif Analyzer that equaled the sum of motif occurrence in the sequence S_i, *i.e.*, $SC_i = \sum_{j=1}^{n} f_{ij}$ (**Fig.3.1b**)

Uncertainty measurement

To characterize the probability of a predicted PRE site being the real PRE site, we performed a non-parametric analysis on 100 randomly generated genomes whose size and nucleotide distribution (A: 29%, C: 21%, T: 29%, G: 21%) are the same as the *D. melanogaster* genome. We used our software to predict PRE sites on these random genomes and for a given score s we counted O_s that denoted the occurrence

of a score that is higher or equal to s. We then calculated E_s, *i.e.*, the E-value of score s by $O_s/100$ and the corresponding P-value would be $E_s/100$.

Genome-wide prediction of PcG target genes

PRE-to-gene Mapper

From our genome-wide PRE site prediction results, the PRE-to-gene Mapper first mapped all of the predicted PRE sites to their genomic coordinates on the genome. When several windows/loci adjacent to each other were all predicted to be PRE sites, they were all combined into a longer PRE. The Mapper then analyzed every locus that had a positive PRE score S, located its most adjacent gene G and credited G a score equaling S. If the locus was positioned closely in between two genes and if the second closest gene G_2 was within 4,000 bps away, the Mapper granted G_2 a score equaling S as well.

Conservation Level Analyzer

Due to their roles in regulating key developmental processes, PcG target genes were expected to be evolutionarily conserved. The Conservation Level Analyzer considered six *Drosophila* genomes that are close to *D. melanogaster* according to the phylogenetic tree [145], including *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. pseudoobscura* and *D. ananassae*. For each annotated *D. melanogaster* gene, it queried the Flybase database (www.flybase.org) to locate its orthologues in any of the six related genomes. If a gene failed to have any orthologue, it was eliminated from

the eligible gene list. That is, the Analyzer excluded the genes that did not have any orthologue in any related species from the pool of candidate genes to be considered as PcG targets. Eventually all the remaining genes were ranked according to the genes' associated PRE scores. The version of our *EpiPredictor* up to this point was termed as *EpiPredictor-Basic.*

Comparative Genomics Analyzer

We investigated the value of incorporating comparative genomics [146-148] into PcG target gene prediction. For this, we constructed a variant version of *EpiPredictor*, hereafter referred to as *EpiPredictor-CG*, which integrated analyses on three well-annotated *Drosophila* organisms (*D. simulans*, *D. yakuba* and *D. pseudoobscura*) that are close to *D. melanogaster* in the phylogenetic tree [145].

Our tactic in implementing *EpiPredictor-CG* was to construct an ensemble system that employed the top-ranked genes provided by our original *EpiPredictor-Basic* as the base set and incorporated the information obtained from our comparative genomics study for rank adjustment when necessary. To be more specific, if our ultimate goal was to retrieve N genes that were most likely the PcG targets, we started our process with a gene list containing the top M genes ranked by *EpiPredictor-Basic* ($M = 1.5N$) and reordered the genes based upon the scores of the candidate genes' orthologues in different *Drosophila* species.

To achieve this, we applied *EpiPredictor-Basic* onto each of these three *Drosophila* genomes. For each genome, all annotated genes were evaluated and ranked according to their predicted PRE scores. If a gene is orthologous to a *D.*

melanogaster gene, the rank of that gene was linked to its *D. melanogaster* gene orthologue. Therefore, for any *D. melanogaster* gene included in the top list, up to four ranks could be obtained, each representing the rank of the gene (or its orthologue) in the respective species, *i.e.*, *D. melanogaster*, *D. simulans*, *D. yakuba* and *D. pseudoobscura*. A final rank was calculated by averaging all the ranks. The gene list was then re-sorted accordingly.

BART-based PRE classifier

BART (Bayesian Additive Regression Trees) is a nonparametric regression method that can also be used as a binary classifier. As a comparison to the SVM-based PRE Classifier in our system, we used BART as an alternative classifier to evaluate whether a given locus is actually the PRE site. This was achieved by using the *R package* (BayesTree) by Hugh Chipman and Robert McCulloch.

Computational complexity

The primary computational complexity of our *EpiPredictor* model came from the component of the SVM-based PRE classifier. During the training phase, the complexity of the SVM was $O(N_s^3 + (N_s^2)l + N_s d_L l)$ where N_s denoted the number of support vectors, l denoted the number of training points and d_L denoted the dimension of input data. During the testing phase, the complexity of the SVM was $O(MN_s)$ where M was $O(d_L)$. In our experiments, $N_s = 21$.

On a regular Dell desktop (Intel Duo CPU 3.00GHz, 1G memory), our system spent 63 milliseconds in training. During the prediction phase, it took about 30 minutes to process the entire *D. melanogaster* genome of 137 million bps and used

around 5 MB memory. Due to the integration of SVM, it was necessary to store a substantial amount of feature vectors onto a text file. This Input/Output process was responsible for the majority of the execution time.

Importantly, our system is an automated program in which the components such as Motif Analyzer, SVM-based PRE Classifier and GC Analyzer were run sequentially requiring no human intervention after the genome sequence under study is input, and is readily scalable. For example, we used our software to predict the PcG target genes on the entire human genome and obtained complete results within three and a half hours on the same PC.

In marked contrast, when we used BART as an alternative to SVM to classify PRE, we noticed that BART required substantial computational resources. It was impossible to complete the prediction of *D. melanogaster* genome on the same PC. On an Intel Xeon computer cluster which contains 134 SunFire x4150 nodes from Sun Microsystems, the computation took about 33 hours to complete. The average usage of memory is 16 GB.

Immunoprecipitation of crosslinked chromatin from *D. melanogaster S2* cells

D. melanogaster Schneider *S2* cells were cultured in 1× Schneider's medium (Invitrogen) supplemented with 20% fetal bovine serum, 100U/ml Penicillin and 100μg/ml Streptomycin at room temperature. Cells were passaged at 1:4 ratio every two days to keep logarithmic growth. Crosslinking, immunoprecipitation with anti-E(z) antibody and quantitative PCR (qPCR) were done as described previously [149]. In brief, 5 μg anti-E(z) (Santa Cruz Biotechnology, Inc) or anti-FLAG mock antibody

(Sigma) were added to 4×10^8 crosslinked *S2* cells to immunoprecipitate protein/DNA complexes. The antibody-protein/DNA complexes were then purified using 50 µl protein A Sephorose 4 Fast Flow Beads (GE Healthcare). DNA was extracted from the purified antibody-protein/DNA complexes by phenol-chlorophorm extraction. Purified DNA was subjected to qPCR using primer pairs designed to amplify DNA of ~250 bps using a SYBER green detection mix (Applied Biosystems). All experiments were carried out in triplicates.

3.2.2 Methods for Human Transcription Factor Discovery

Computational motif discovery

22 different CHIP data sets (**Table 3.2**), using the measured targets for either promoters (CHIP-CHIP) or specific loci (CHIP-seq), had their sequences downloaded from the Ensembl genome construct, and were applied to the five motif discovery algorithms [8, 150-152]. The top three motifs from each run were collected and converted into position weight matrixes (PWM) for further analysis. Each program was run using its default settings on a 16-node clustered CPU. To analyze the effect of genetic conservation, the same data sets were modified to only include sequences that were qualified as conserved against a 31 eutherian mammals GERP score as defined within the Ensembl genome browser. After shortening all sequences to only the conserved elements, the new data set was then reapplied to the five motif discovery algorithms and their top three motifs were collected and converted to PWMs. The output motifs for these programs were named as a combination of the

program name, which experimental data set it came from, whether it was from a conserved sequence alignment (as discussed below), and its motif ranking amongst the output of that run.

Table 3.2. List of all 22 cell lines and antibodies used for CHIP experiments from literature.

Cell line	Antibody used	Reference
hES	H3K27	[153]
H1	H3K27	[81]
H9	Suz12	[78]
H9	EED	[78]
H9	H3K27	[78]
H9	H3K27	[82]
Ntera2	Suz12	[72]
PC3	H3K27	[154]
SW480	Suz12	[72]
MCF7	Suz12	[72]
TIG3	Suz12	[77]
TIG3	H3K27	[77]
Gastric adenocarcinoma	H3K27	[116]
RL	Suz12	[155]
RL	H3K27	[155]
EP156T	H3K27	[87]
EP156T	H3K27	[156]
EPT1	H3K27	[156]
EPT2	H3K27	[156]
Dendritic cells	H3K27	[157]
Macrophages	H3K27	[157]
Monocytes	H3K27	[157]

For example, the motif 7ggm2 ie. (7)(g)(gm)(2) was generated from the data present in run number 7, using only conserved sequences (g) and the program (gm) GADEM, and was the second highest scoring of that batch (2). Additional

abbreviations are: md (MDscan), ace (AlignACE), and meme (MEME). Another set of suffixes refers to the stringency of the CisGenome run setting while mapping the PWM to the genome. These are no suffix, in which default run settings are used; -high, where the Likelihood Ratio setting is set to 2000; -con, where the filter by conservation tab is checked and set to being CS>=66; and –highcon, where the two filters just mentioned are used together at the same time.

Additionally a set of GADEM and CisGenome motif discovery runs were taken against CpG or LcG specific regions independently of any CHIP data. These are labled Lcg_# for LcG specific; lcg3+-# for promoters having at least 3 CpG islands in their promoter, while still falling in the LcG list; and cpg_edge, for using regions that extend +- 250 bp from defined CpG island edges, as PRC2 peaks often appear in these regions. Motif1-6 are a special class of GADEM run motifs which were computed against a composite list of promoters where the presence or absence of CpG islands was explored. These motifs showed to be specific to either: promoters with CpG islands in them(motif1 and motif6), ES promoters with CpG islands in them(motif4), ES without CpG islands in them(motif2), or ES and Cancer promoters regardless of CpG content(motif3 and motif5).

Motif analysis

The motif discovery algorithms collectively returned a list of 65 motifs to be analyzed. To compare each motifs' overall fitness, we mapped these motifs to a 6000 bp window representing positions -5000 to +1000 with respect to the TSS was downloaded from the UCSC genome browser (hg19) for all genes in the human

genome using CisGenome. It was to be expected that many of these motifs might be redundant, so to address this, if any two motifs were found to co-occupy a given 5 bp sliding window more than 75% of the time, they were assumed to be variations of the same motif, and the smaller of the two motifs was eliminated from the list of potential motifs. This left 39 motifs. An enrichment score was then generated by dividing the number of genes whose promoter region contained a given motif from any of the data sets of interest with an appropriately sized set of randomly selected genes. The same methodology was carried out for PWM's from the Jaspar or Transfac database of human transcription factor binding sites and applied them to these same gene lists.

Model Prediction

To determine the extent that these motifs could predict the presence or absence of genes found in the various CHIP experiments, a simple linear addition model was chosen in which the presence or absence of a motif is represented as a binary function and then multiplied by a binary modifier function to either allow or disallow a motif in the matrix from being considered. This value is then summed across every promoter in the genome creating a score ranging from 0(has no motifs) to 331(has every motif). These scores are then ranked and compared against the known CHIP gene lists by computing AUC scores using ROC plot. The Modifier function is then iteratively modified to include only the motifs that meet different thresholds of enrichment, until a peak is found. The highest scoring AUC score reported here used 155 motifs with enrichments greater than 1.2 according to the

NTerra2 cell line enrichments as applied to the H9suz12 CHIP list of 1040 genes. It was not our intention to find the best model prediction computationally possible, only to demonstrate that our motifs have predictive power, supporting the idea that DNA elements should be functionally important towards PRE discovery.

3.3 Genome-wide Polycomb Target Gene Prediction in *Drosophila melanogaster*

3.3.1 Results

Empirical analysis of the SVM-based PRE classifier

To identify the most appropriate kernel for the SVM-based PRE classifier, we performed an empirical analysis on the training set to gauge how well a certain kernel distinguished known PRE sequences. This is done using three runs of 10-fold cross validation so as to avoid any potential over-fitting problem. With the default parameters provided by LibSVM [158], the performance of all four basic kernel methods was analyzed by sensitivity and specificity (**Table 3.3**). As we expected, the non-linear kernels such as polynomial worked very well in distinguishing PRE sequences from control sequences, further confirming the advantage of modelling the motif interaction in a combinatorial manner. Among them, the polynomial (d=2 and d=3) kernels (also called the *quadratic kernel* and *cubic kernel*, respectively) achieved the best results in terms of specificity and sensitivity when both the average and standard deviation are taken into account, implicating that at the PRE sites,

multiple transcription factors interact with each other that as a whole serves as the platform for PcG recruitment. Although the cubic kernel did not significantly outperform the quadratic kernel, it is still the best model given all the parameters considered. Therefore, we used the cubic kernel on the SVM throughout our analyses.

Table 3.3. SVM kernel evaluation

Kernel / **Metric**	*Linear*	*Polynomial (d=2)*	*Polynomial (d=3)*	*RBF*
Sensitivity	0.80±0.05	0.80±0.05	**0.82±0.03**	0.60±0.05
Specificity	0.91±0.01	0.96±0.01	**0.96±0.01**	0.99±0.02

Sensitivity = TP/(TP + FN); Specificity = TN/(TN + FP),
where TP, TN, FP, FN correspond to true positive, true negative, false positive and false negative, respectively. We performed three independent runs of 10-fold cross validation on the training collection and reported the average sensitivity/specificity and the standard deviation.

Test of the training set

To compare our new sequence collection with the original one used by Ringrose and colleagues, we ran independently *EpiPredictor-Basic* using the modules (a,b,c) on both training sets and the *jPREdictor* (static) on our new training set (**Table 3.4**). Also included in Table 3.4 is the result of *jPREdictor* (static) with Ringrose's training set as originally reported [125]. We found virtually no difference in performance when using different training sets. Therefore, we elected to use our training set throughout the analyses.

Table 3.4. Comparison of our new training set (New) with Ringrose's original training set (Ringrose)

Methods	Training sets	Schwartz *et al.* [1]	Tolhuis *et al.* [2]	Schuettengruber *et al.* [3]
EpiPredictor-Basic (a,b,c)	Ringrose	26.14%	10.22%	25.12%
	New	26.14%	10.22%	25.12%
jPREdictor (static)	Ringrose[4]	21.02%	8.00%	20.00%
	New	21.02%	8.00%	21.40%

(a): Motif Analyzer
(b): SVM Classifier
(c): GC Analyzer
[1]: Overlap with the genes predicted by Schwartz *et al.* [75]
[2]: Overlap with the genes predicted by Tolhuis *et al.* [86]
[3]: Overlap with the genes predicted by Schuettengruber *et al.* [85]
[4]: Data reported in the original publication [125]

Performance evaluation of *EpiPredictor* components

We tested our classifier on the *D. melanogaster* genome that contains roughly 137 million bps and 13,000 genes. Each chromosome was scanned with a sliding window of 200 bps and a step size of 20 bps (parameters determined by empirical analysis), and each window was analyzed by the Motif Analyzer component and represented by a seven-dimensional feature vector (each corresponding to one of the seven motifs we used). The performance of the system was evaluated by the matching ratios between our top predicted genes and those of the three validation sets derived from ChIP studies [75, 85-86] together with their intersection set (Intersection) (**Table 3.6**). To examine whether the performance of our system is sensitive to different window size and step sizes, we also varied the values of these parameters (**Table 3.5**). It is clear that with different window and step sizes, the performance of our system did vary slightly but the change was not very substantial.

Overall, the parameter setting of window size = 200, step size = 20 produced the best results. Therefore we used the window size of 200 bps and step size of 20 bps throughout.

Table 3.5. Performance analysis of different window sizes and step sizes in Motif Analyzer

Window/Step Size	Schwartz *et al.* [1]	Tolhuis *et al.* [2]	Schuettengruber *et al.* [3]	Intersection [4]
W=200, S=20	26.14%	10.22%	25.12%	23.68%
W=300, S=30	26.70%	8.89%	25.12%	21.05%
W=400, S=40	25.57%	8.44%	24.19%	21.05%
W=500, S=50	24.43%	8.44%	22.79%	23.68%
W=500, S=10	25.57%	10.22%	24.19%	15.79%

All of the percentages shown above indicate the overlap between the top 243 predicted genes using the *EpiPredictor* components (a,b,c) and those predicted by the ChIP studies.
[1]: Overlap the genes predicted by Schwartz *et al.* [75]
[2]: Overlap with the genes predicted by Tolhuis *et al.* [86]
[3]: Overlap with the genes predicted by Schuettengruber *et al.* [85]
[4]: Overlap with the genes intersected by Schwartz *et al.*, Tolhuis *et al.*, and Schuettengruber *et al.*

Our system contains multiple components. The effect of each component was evaluated by sequentially adding each component onto the Baseline system that used only the Motif Analyzer.

Table 3.6. Evaluation of the performance of individual *EpiPredictor* components against three genome-wide ChIP studies in D. melanogaster and their intersection

Number of Top Genes	*EpiPredictor* Components	Schwartz *et al.* [5]	Tolhuis *et al.* [6]	Schuettengruber *et al.* [7]	Intersection [8]
243 [1]	*(a)*	14.20% [9]	5.33%	12.09%	2.63%
	(a,b)	22.73%	9.78%	19.53%	23.68%
	(a,b,c)	26.14%	10.22%	25.12%	23.68%
	(a,b,c,d)[3]	**27.27%**	**10.67%**	**26.05%**	**26.32%**
322 [2]	*(a,b,c,d)*	32.39%	14.22%	30.70%	34.21%
	(a,b,c,d,e)[4]	**35.80%**	**15.11%**	**33.02%**	**44.74%**

(a): Motif Analyzer
(b): SVM Classifier
(c): GC Analyzer
(d): Conservation Level Analyzer
(e): Comparative Genomics Analyzer
[1]: The number of top genes retrieved from *EpiPredictor-Basic* analysis.
[2]: The number of top genes retrieved from *EpiPredictor-CG* analysis.
[3]: The *EpiPredictor-Basic* module.
[4]: The *EpiPredictor-CG* module.
[5]: Overlap with the genes predicted by Schwartz *et al.* [75]
[6]: Overlap with the genes predicted by Tolhuis *et al.* [86]
[7]: Overlap with the genes predicted by Schuettengruber *et al.* [85]
[8]: Overlap with the genes intersected by Schwartz *et al.*, Tolhuis *et al.*, and Schuettengruber *et al.*
[9]: Suppose the validation set includes *V* genes. Among the top *N* genes predicted by our system, *W* genes matched the validation set, the overlap was represented as *W/V*.

Baseline system To thoroughly evaluate the merit of each component of *EpiPredictor*, we constructed a baseline system that did not incorporate SVM or any other subsequent component but instead only used the Motif Analyzer that calculated the sum of the motif occurrence frequency. The baseline system achieved a moderate performance, having the matching ratios of 14.20%, 5.33%, 12.09%, 2.63%, with the

three validation sets and their intersection, respectively (**Table 3.6**). It is noteworthy that to perform a fair comparison with *jPREdictor* that reported their top 243 genes, we also retrieved the top 243 genes from our system to obtain the aforementioned results.

SVM-based PRE Classifier To estimate the merit of SVM, we then integrated SVM into the baseline system. The application of SVM drastically enhanced the performance of our system when compared to the baseline system, with matching ratios of 22.73%, 9.78%, 19.53%, 23.68%, respectively (**Table 3.6**).

GC Analyzer Subsequently we incorporated the GC Analyzer into our program. The prediction performance of *EpiPredictor* was further improved to 26.14%, 10.22%, 25.12%, 23.68%, respectively (**Table 3.6**), demonstrating that the *bona fide* PcG sites tend to have relatively high GC content.

Uncertainty measurement The non-parametric tests conducted on 100 random genomes indicated that a PRE score of 12.7 corresponded to a *P*-value of 0.01. In our prediction, the top 190 predicted PRE sites had a PRE score of higher than 12.7, with the highest score being 39.2. We also corrected the issue of multiple comparisons using Bonferroni correction, and found that a PRE score of 17.3 corresponded to a *P*-value of 0.0001 (0.01/100). In our prediction, the top 73 predicted PRE sites had a *PRE* score of 17.3 or higher. Thus these top 73 predicted PRE sites are regarded as predictions with significant confidence, even under such a stringent condition.

Conservation Level Analyzer The integration of the Conservation Level Analyzer slightly enhanced our system's performance to 27.27%, 10.67%, 26.05%,

26.32%, respectively (**Table 3.6**). At this point, we completed the construction of the basic version of our system, *EpiPredictor-Basic*. A complete list of the top genes thereby generated is provided in Appendix Table 4.2.

It is worth mentioning that an attempt of using the base-by-base evolutionary conservation score compiled on *D. melanogaster* genome in comparison to 14 insects [148] failed to produce any improvement in the prediction performance (data not shown). Taken together, this suggested that the *bona fide* PcG target genes be most likely evolutionarily conserved; however, their positions might be more flexible in the course of evolution.

Comparative Genomics Analyzer To evaluate the performance of *EpiPredictor-CG*, which integrated the Comparative Genomics Analyzer, we retrieved the top 322 predicted genes, which was the same number as generated by our counterpart *jPREdictor* (dynamic) (**Appendix Table 4.3**). Due to the integration of comparative genomics, some of the genes with lower scores were boosted up into the top list and yielded an improved performance of 35.80%, 15.11%, 33.02%, 44.74%, respectively (**Table 3.6**), in comparison to the performance of *EpiPredictor-Basic* in predicting 322 genes: 32.39%, 14.22%, 30.70%, 34.21% (**Table 3.6**).

It is worth noting that the Intersection set obtained by intersecting all the three validation sets derived from ChIP studies [75, 85-86] did have very high matching ratio with our *EpiPredictor-CG* prediction (**Table 3.6**), consistent with the expectation that it is the highest confidence set of the target genes.

Performance comparison between SVM-based and BART-based PRE classifier

Besides SVM, several other statistical models including BART [159] are also able to capture nonlinear interactions among the sequence features. For instance, Liu *et al.* used BART to predict polycomb target genes with a good performance [126]. Therefore we compared our system's performance using SVM-based or BART-based PRE classifier (**Table 3.7**). It is clear that the SVM-based classifier consistently outperformed the BART-based counterpart.

Table 3.7. Evaluation of the performance of our system using SVM-based PRE classifier vs BART-based PRE classifier

Method	*EpiPredictor* Components	Schwartz *et al.*[1]	Tolhuis *et al.*[2]	Schuettgurber *et al.*[3]	Intersection[4]
SVM	(a, b)	**22.73%**	**9.78%**	**19.53%**	**23.68%**
	(a, b, c)	**26.14%**	**10.22%**	**25.12%**	**23.68%**
BART	(a, d)	21.59%	8.44%	19.07%	21.05%
	(a, d, c)	22.73%	9.33%	22.79%	21.05%

(a): Motif Analyzer
(b): SVM-based Classifier
(c): GC Analyzer
(d): BART-based Classifier
[1]: Overlap between the top 243 predicted genes with the genes predicted by Schwartz *et al.* [75]
[2]: Overlap between the top 243 predicted genes with the genes predicted by Tolhuis *et al.* [86]
[3]: Overlap between the top 243 predicted genes with the genes predicted by Schuettengruber *et al.* [85]
[4]: Overlap between the 243 predicted genes with those intersected by Schwartz *et al.*, Tolhuis *et al.*, and Schuettengruber *et al.*

Comparative analysis of *EpiPredictor* and *jPREdictor*

We conducted a comparative analysis of *EpiPredictor* and *jPREdictor* (**Table 3.8**) by using the matching ratios as well as the receiving operating characteristics (ROC) curve as our evaluation metrics. The former metric indicates the overall accuracy of prediction while the latter one depicts the trade-off between sensitivity and specificity, which focuses on evaluating the ranking scheme. In terms of the

matching ratio, *EpiPredictor-Basic* outperformed *jPREdictor* (static) by 6.25%, 2.67%, 6.05%, 5.27%, respectively, against the three validation sets and their intersection set and the improvement is statistically significant ($P<0.05$ in one-tailed Students' *t*-test). In addition, *EpiPredictor-CG* surpassed the performance of *jPREdictor* (dynamic) by 7.96%, 2.67%, 10.23%, 18.42%, respectively ($P<0.05$). In terms of the area under curve (AUC) of ROC curve, *EpiPredictor-Basic* achieved comparable results with *jPREdictor* (static) whereas *EpiPredictor-CG* outperformed *jPREdictor* (dynamic) in three out of the four cases (**Fig.3.2**). It is worth noting that the AUCs of *EpiPredictor-Basic*, *EpiPredictor-CG* and *jPREdictor* (static) were all significantly larger than 0.5 (random guess) ($P < 0.05$) but it was not the case for *jPREdictor* (dynamic). Furthermore, using the AUCs as a measure, neither *EpiPredictor* nor *jPREdictor*'s advanced version significantly outperformed their basic counterpart.

Table 3.8. Comparison of the overlaps between the PRE genes predicted by *EpiPredictor* and *jPREdictor* and three genome-wide ChIP studies in D. melanogaster and their intersection

Scheme	Approach	Schwartz *et al.*[1]	Tolhuis *et al.*[2]	Schuettengruber *et al.*[3]	Intersection[4]
Original (243 genes)	*EpiPredictor-Basic*	**27.27%**	**10.67%**	**26.05%**	**26.32%**
	jPREdictor (static)[5]	21.02%	8.00%	20.00%	21.05%
Comparative Genomics (322 genes)	*EpiPredictor-CG*	**35.80%**	**15.11%**	**33.02%**	**44.74%**
	jPREdictor (dynamic)[5]	27.84%	12.44%	22.79%	26.32%

[1]: Overlap with the genes detected by Schwartz *et al.* [75]
[2]: Overlap with the genes detected by Tolhuis *et al.* [86]
[3]: Overlap with the genes detected by Schuettengruber *et al.* [85]
[4]: Overlap with the genes intersected by Schwartz *et al.*, Tolhuis *et al.* Schuettengruber *et al.*
[5]: Data reported in the original publication [125]

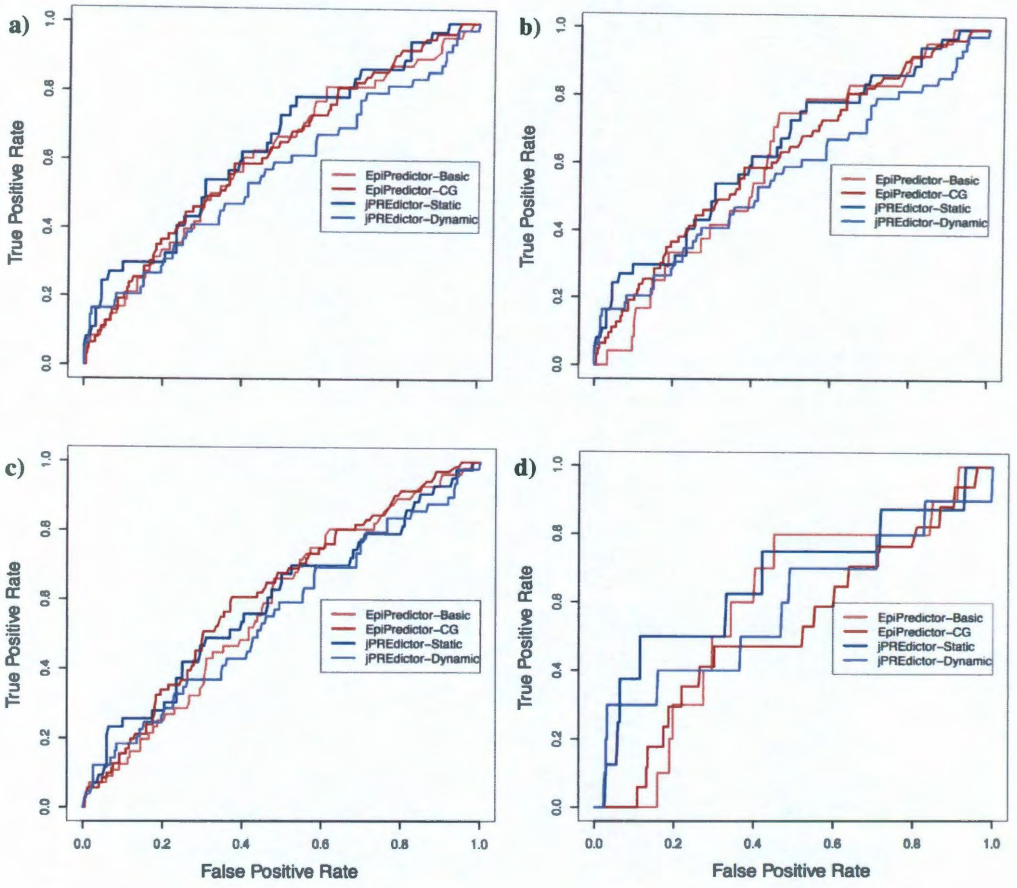


Fig.3.2. ROC curves of the PRE genes predicted by *EpiPredictor* and *jPREdictor*. Shown are overlaps with the genes predicted by Schwartz *et al.* (A), Tolhuis *et al.* (B), Schuettengruber *et al.* (C) and the genes intersected by all three sets (D). The AUCs on the four validation sets are 0.61, 0.61, 0.58 and 0.60, respectively, for *EpiPredictor-Basic*, 0.62, 0.57, 0.62 and 0.53, respectively, for *EpiPredictor-CG*, 0.64, 0.56, 0.59 and 0.67 for *jPREdictor* (static), 0.56, 0.49, 0.55 and 0.59 for *jPREdictor* (dynamic).

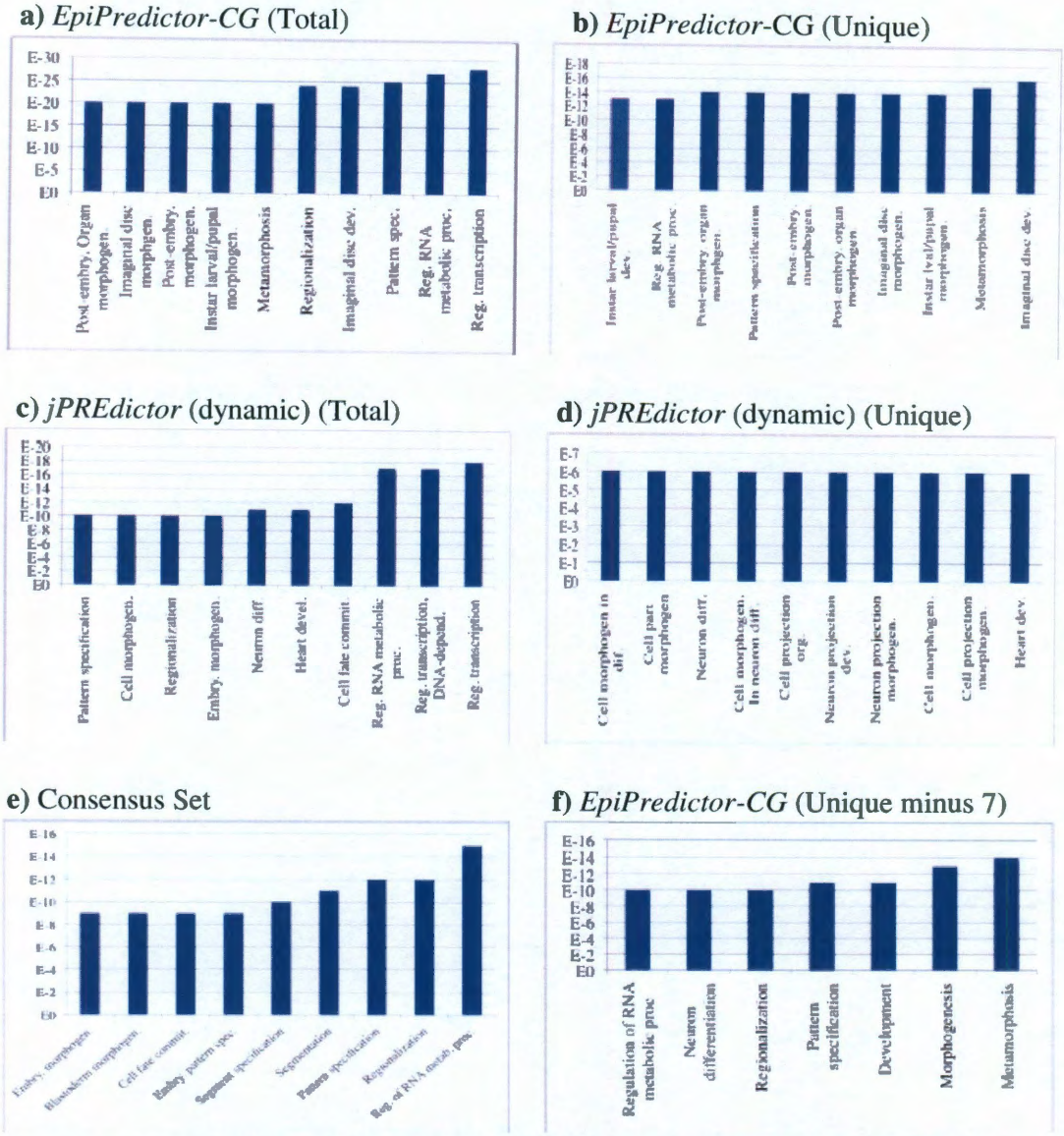


Fig.3.3. Gene ontology analysis of genes predicted by *EpiPredictor* and *jPREdictor*. Shown are the top ten gene ontology terms related to the genes predicted by: A) *EpiPredictor-CG*; B) *EpiPredictor-CG* but not *jPREdictor* (dynamic); C) *jPREdictor* (dynamic); D) *jPREdictor* (dynamic) but not *EpiPredictor-CG*; E) both EpiPredictor-CG and *jPREdictor* (dynamic); F) *EpiPredictor*-CG except the seven annotated genes.

Annotation of *EpiPredictor* prediction

To reveal the major function enrichment of the genes predicted by our system and *jPREdictor*, we used the DAVID bioinformatics tool [160] to perform a gene ontology analysis on the genes uniquely predicted by either *EpiPredictor-CG* or *jPREdictor* (dynamic), as well as those predicted by both *EpiPredictor-CG* and *jPREdictor* (dynamic) (**Fig.3.3**). Most of the highly represented gene functions were related to the regulation of transcription, development, pattern specification, morphogenesis, and cell fate commitment, consistent with the expected roles of PcG in regulating key developmental processes of an organism [75, 77-79, 84-87]. The consensus genes predicted by both *EpiPredictor-CG* and *jPREdictor* (dynamic) made up about 28% of the top 322 genes and their corresponding gene ontology analysis presented good consistency with experimental studies.

By cross-referencing existing literature, we found experimental evidence for seven genes, which were uniquely identified by *EpiPredictor-CG* and also matched at least one of the three ChIP studies, of their critical roles in key developmental processes (**Table 3.9**). To exemplify, the *inv* locus was recently found to harbor one PRE site which has been experimentally verified [111] and its role in regulating *Drosophila* hindgut development is well established [161]. The *wg* locus belongs to the important Wg/Wnt signal transduction pathway that directs a variety of cell fate decisions in developing animal embryos[162]. In *Drosophila*, *wg* alone directs a wide range of cell fate and patterning decisions [163]. The *nub* locus is involved in embryogenesis and neurogenesis [164-166]. The *pdm2* locus is responsible for a

variety of cell fate decision in the *Drosophila* development [167]. The *dac* is an essential part of a complex that functions to induce ectopic eye development [168]. The *Gsc* mediates effective repression in *Drosophila* blastoderm embryos [169]. The *tup* has a key function in the development of imaginal disc [170] and is also a key component in early cardiogenesis [171]. Interestingly, a recent ChIP study [78] revealed that the human homologues of *wg* (WNT1), *dac* (DACH-1), *Gsc* (GSC) *and tup* (ISL1) are all targeted by PcG. In particular, WNT1 is known to be involved in embryogenesis and cancer development [172]. The functions of the genes uniquely identified by our system but excluding the abovementioned seven genes are shown by a gene ontology analysis using DAVID (**Fig.3.3F**).

Table 3.9. Annotation of a set of seven genes uniquely identified by *EpiPredictor*

Gene	Verified Function in *Drosophila*	Vertebrate Homologue
inv	A newly experimentally validated PRE was found to exist in the *inv* locus [111]. It is important for hindgut development [161]	
wg	Embryogenesis (Wingless/Wnt signaling pathway) [163]	WNT1: predicted as PcG target in human [78]; involved in embryogenesis and cancer [172]
nub	Embryogenesis, neurogenesis [164-166]	
pdm2	Important for a variety of cell fate decisions in development [167]	
dac	Induce ectopic eye development [168]	DACH-1: predicted as PcG target in human [78]
Gsc	*groucho*-dependent repression in embryo [169]	GSC: predicted as PcG target in human [78]
tup	Imaginal disc development [170], key component in early cardiogenesis [171]	ISL1: predicted as PcG target in human [78]

To further validate our prediction, we also cross-referenced our gene list with the 27 PcG target genes confirmed by ChIP-qPCR in the work of Ringrose and colleagues [113], of which *EpiPredictor-CG* correctly predicted 19 genes (**70%**), exhibiting a good correlation.

Experimental validation of *EpiPredictor* prediction

In order to experimentally validate *EpiPredictor* prediction, ChIP-qPCR was used to investigate the enrichment of 15 predicted PRE sites that were randomly selected from the top 150 predictions (**Appendix Table 4.4a**) using anti-E(z) antibody. For positive controls we used three known PREs, *bxd, iab2,* and *en_DM*, as established in the literature [173] along with four sequences from Ringrose *et al* [113], *hth, unc-4, idgf4,* and *cato*, for which ChIP-qPCR experiments have been done using anti-PC antibody. Three housekeeping genes with no previous evidence as PRE or of polycomb related activity, *hsp22, hsp26*, and *Pc*, were selected as negative controls (**Appendix Table 4.4b**).

The results of ChIP-qPCR showed that there are more than two-fold enrichments for 12 out of the 15 tested PRE sites (**Fig.3.4**). Among them, five showed enrichment greater than the average value of 5.66 for the seven positive controls, indicating a higher degree of confidence for their potential as PcG target genes. Our E(z)-ChIP derived data and Ringrose's PC data are scaled roughly to the same level (**Table 3.10**) with the exception of idgf4 which exhibited enrichment in our data but not in Ringrose's [113]. However, this discrepancy is not completely

unexpected given the fact that on the whole genome scale PC and E(z) do not always align well [75].

Table 3.10. Comparison of the qPCR data of anti-E(z) ChIP and anti-PC ChIP

	E(z)	PC
hth	6.97	11.9
unc-4	5.42	8
ldgf4	5.5	0.6
cato	0.64	0.5

By mapping the positively enriched sequences onto their closest genes, we found that all 12 corresponding genes are of crucial importance to *Drosophila* embryonic development, since the knockout of each of these genes conferred serious body morphological changes. The *antp* and *abd-A* are *Drosophila* HOX genes [174], while *bxd* is expressed directly upstream of and is known to directly influence the behavior of *ubx*, another HOX gene [175]. Furthermore, both *disco* and *eve* regulate the localization or expression of HOX genes, conversely, *salm* and *bab2* are directly regulated by HOX genes [176-179], while *unc-4* is a homeobox-containing protein and a paralogue of the HOX genes with similar functions [180]. Finally, both *noc* and *pnr* are critical for proper eye formation [181-182], *grn* has importance in multiple organ development [183] and immune response in the midgut [184], and *zfh1* is essential to cell differentiation of lateral mesodermal derivative lineages and in neurogenesis [185]. The critical importance of these genes and the computational prediction of them being PcG target genes highlight the importance of understanding how sequence influences PcG binding in order to properly understand embryonic development in *Drosophila*.

PcG target genes are essentially free of transposons

Transposons are mobile genetic elements that can cause mutations and change the amount of DNA in the genome [186]. Given their critical importance in cellular functions, we predicted that PcG target genes in *D. melanogaster* should have a minimal presence of transposons, generally termed as transposon-free regions (TFRs). We performed a whole-genome search and identified 1,400 TFRs of >10,000 bps in length, of which 1,232 overlapped with at least one gene's TSS. In the top 322 putative PcG target genes predicted by our system, 319 of them (99%) had TSS overlapping with one or more TFRs. Thus, as we expected, the *D. melanogaster* PcG target genes are indeed essentially free of identifiable transposon-derived sequences. This is a novel finding in *Drosophila* and corroborates well with several recent mammalian studies that revealed strong correlations between TFRs and genes encoding developmental regulators [187], as well as the H3K27me3 marks [188].

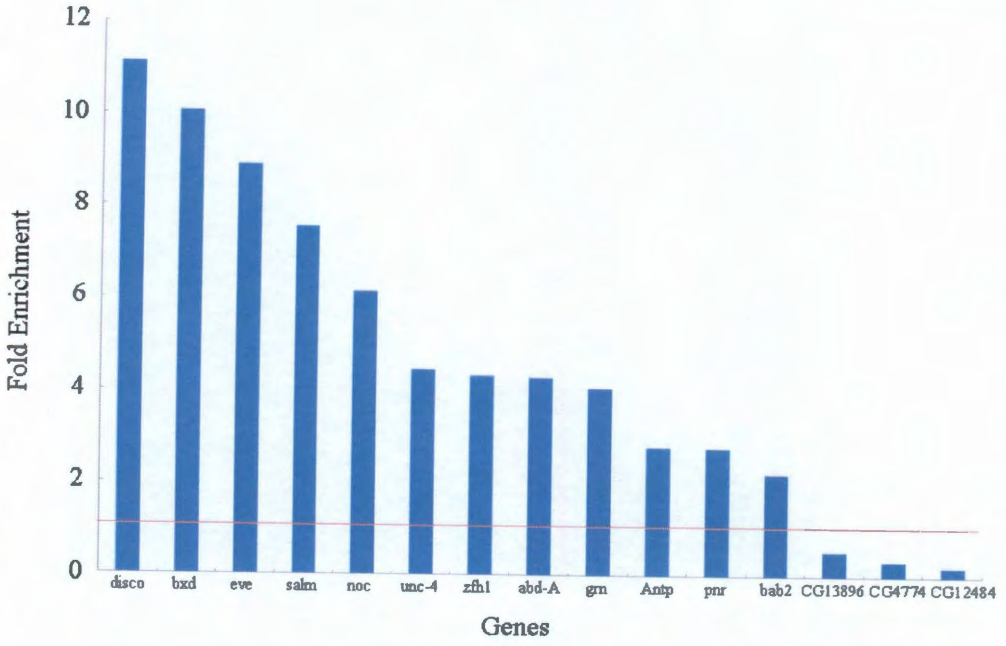


Fig.3.4. ChIP-qPCR verification of *EpiPredictor* prediction.
Shown are the enrichment of each genomic region (predicted PRE site) in *S2* cell ChIP samples using anti-E(z) versus the use of anti-FLAG mock antibodies. The horizontal line shows an enrichment of 1 (no enrichment). The gene symbols listed are those of the genes closest to the tested genomic regions. For specific coordinates please refer to Appendix Table 4.4.

3.3.2 Discussion

Sequence ambiguity and multi-motifs in *EpiPredictor*

Given the ambiguity in the consensus sequence of motifs, our system considered different versions of the same motifs (for instance, PS, PM, and PF for Pho) as well as allowed the existence of ambiguity codes and mutations (**Table 3.1**). In addition, by using an SVM with non-linear kernel as a PRE classifier, our program abstractly models how multiple motifs interact with each other at the genomic site of interest. These two considerations are similar to the options of position-specific probability matrices and multi-motifs in *jPREdictor*.

Transcription factor networking is important for PcG recruitment

To the best of our knowledge, this is the first application of SVM to PRE prediction. With the integration of a non-linear kernel, our system *EpiPredictor* succeeded in modeling the spatial relationship and combinatorial interaction among transcription factors that are involved in PcG recruitment. This strategy offers a higher level of abstraction over any other approaches that use a linear function. The fully automated process of constructing the classifier in SVM also reduces the level of bias in the analysis.

Our novel computational strategy also offers new insights into the interactions among transcription factors at the *cis*-regulatory elements *in vivo*. The outstanding performance of the non-linear kernels indicates that multiple transcription factors are networking at the *cis*-regulatory elements for efficient recruitment of PcG proteins. However, the details of such networking remain to be illustrated in future studies.

High GC content and conservation level are important features of PcG target genes

Among the array of perspectives that we used in *EpiPredictor*, SVM classifier, high GC content and comparative genomics all led to substantial performance improvements (**Table 3.6**). The success of integrating GC analysis suggested that relatively high GC content be an important feature of PcG target genes, consistent with previous studies that hyper-conserved CpG domains underlie polycomb-binding sites [142]. In addition, given their critical importance in cellular functions, PcG target genes are not surprisingly highly conserved in evolution.

PcG target genes are essential for transcription and development

The gene ontology analysis on the genes predicted by our system revealed that the target genes of PcG are mainly regulators of transcription activities and are crucial for key developmental processes. Some genes uniquely predicted by our system are confirmed by several independent experimental studies to be essential for normal development and patterning. These observations further support the fundamental roles of PcG proteins in development and cellular functions.

Prediction of TrxG target genes

Trithorax group (TrxG) proteins methylate histone 3 lysine 4 to reverse the repression imposed by PcG proteins [90, 189]. There exists substantial evidence that Trithorax response elements (TREs) and PREs co-localize. For example, several major TrxG proteins bind at essentially all known or presumptive PREs, suggesting that the regulatory platforms are switchable [90, 190]. In mouse embryonic stem cells, large bivalent domains were found to contain chromatin modifications generated by both PcG and TrxG, suggesting the co-presence of PcG and TrxG in developmental genes [188]. A recent genetic study on *Drosophila* also revealed that PcG repression is dynamic and that ASH1 (absent, small or homeotic discs 1), the histone methyltransferase belonging to the TrxG complex, is critical for the active state of Polycomb target genes [190]. Taken together, accumulating evidence suggests that the epigenetic regulations mediated by PcG and TrxG are likely to be closely intertwined and that the approach that accurately predicts PcG target genes will also shed new light on TrxG target genes. Thus it would not be surprising if the PcG target genes we predicted here will turn out to be TrxG target genes as well.

Concluding Remarks

Despite a large number of genome-wide ChIP studies of PcG target genes [72, 75-79, 81, 83, 85-87] recently appeared in the literature that substantially enriched our knowledge of the scales of PcG-mediated epigenetic modification and their roles in normal cellular functions and in cancer development, our mechanistic understanding of this process remains extremely poor. To exemplify, up to date, there are only two mammalian PREs [80, 114] and a dozen of *Drosophila* PREs [104-113] that have been experimentally verified. In addition, there are only nine *Drosophila* transcription factors confirmed to be involved in PcG recruitment, among which only two have mammalian homologues [92, 191]. The extremely limited pools of confirmed PREs and their interacting transcription factors are the main restraints for the relatively mediocre performance of computational methods such as *EpiPredictor* and *jPREdictor*, with 20~30 % matching ratios with genome-wide ChIP data. Although our *EpiPredictor* has substantially outperformed *jPREdictor* (by up to >10% in matching ratio), we expect a much better performance if we had had more knowledge on PREs and their interacting transcription factors. Thus, the more accurate computational method such as *EpiPredictor* will provide a very useful tool for initial screening of PcG target genes from ChIP studies so as to identify the most likely candidates for labor-intensive experimental verifications. The enhanced knowledge of PREs will in turn improve the performance of these computational methods, and ultimately leads to a comprehensive understanding of PcG-mediated gene repression in normal cellular functions as well as in epigenetic dysregulation.

Thus, our new *EpiPredictor* program reported in this study represents an important step toward this ultimate goal in the field of epigenetics.

3.3 Computational Discovery of DNA Motifs Specifying H3K27-Mediated Gene Silencing

3.4.1 Results

Due to the heterogeneous nature of the data, which resulted in varying definitions for what genes were looked at and what constituted a promoter, a consensus region was selected consisting of -5000 bp upstream and +1000 bp downstream with respect to the transcriptional start site (TSS) of the human build hg19 gene annotations [192]. It was then from this promoter region that all subsequent computational experimentation was conducted.

Looking across all published CHIP experiments against PRC2 and/or H3K27me3, a consensus list of 758 promoters (**Appendix Table 4.5**) that have a consistently high CHIP intensity score (as defined by the level of significance in the CHIP signals as originally reported) while still being present in at least 50% of all experiments. These genes, having been found to be consistently modified across a diverse set of developmental stages representing Embryonic Stem Cells (ESC's), a variety of cancer types, and normal differentiated epithelial cell lineages, thus represent the most comprehensive list of potential PRE's to date.

Computational Identification of Conserved DNA Motifs

To identify conserved DNA motifs that are enriched in genes identified in CHIP experiments against PRC2 and/or H3K27me3, we ran multiple motif finding algorithms, MEME [193], MDscan [194], AlignACE [195] (which provided no unique motifs), Gadem [196], and the Gibbs Motif Sampler from CisGenome [197]. The results of this motif search identified more than 100 different motifs. After combining or removing very similar motifs, the final list contained 50 unique motifs (**Table 3.11**).

To further characterize the identified motifs, we calculated the occurrence of these unique motifs across a variety of gene categories of interest including: the top 758 identified above, as well as their specific enrichment against all of the 22 CHIP gene lists, separately, and against the cellular state (ESC, cancer, and differentiated) as an average, and also against known promoter features, such as, CpG Island presence, conservation, and against the expression profile of many select cancer tissues and cell lines using the Oncomine database[129].

These enrichments are then compared against various lists of between 500 to 2000 randomly selected "background" genes not found in any CHIP data. The enrichment of those motifs was represented as the ratio of the occurrence among these lists as compared to what would be expected from a random list of the same size. The motifs with enrichment score of higher than 1.5 for its category are reported in Appendix Table 4.6. For comparison these are slightly higher than the enrichment scores reported in Ku et al, 2008, the highest of these being 7gmd-highcon, which is found in 1936 promoters and has an enrichment of 2.63 against the

Table 3.11. 50 unique computationally derived motifs and the programs that were used to create them. Shown on the left of each motif list is the number of genes in the genome which contain this motif within their promoter.

GADEM		Mdscan		MEME		CisGenome	
4167	lcg3+_gad_3	9380	17md	1086	17gmeme	3073	mat4-1cons
2384	lcg3+_gad_8	10715	7gmd			3961	cpg_edge-1-1
12521	lcg3+_gad_6	14775	10md			203	mat_0cons
12191	lcg3+_gad_1	6557	14md			671	lcg3+-5
11299	lcg3+_gad_7	19793	7gmd2			5422	cpg_edge-3-2
1896	lcg3+_gad_5	18595	1gmd2			11029	Lcg_8
11343	motif1	18894	1gmd			13791	lcg3+-4
14911	motif2					14758	Lcg_5
12826	ctrl2					12305	lcg3+-2
11908	ctrl1					8071	cpg_edge-2-1
14927	ctrl3					15199	Lcg_2
11127	motif6					9424	Lcg_3
14585	motif4					9531	lcg3+-9
16645	10gm3					10124	lcg3+-1
16659	10gm2					12903	Lcg_7
16712	lcg_withcpg_gad_3					12144	758m3
15547	motif3					11702	lcg3+-6
17678	17ggm4					9405	lcg3+-0
18334	motif5					1011	cpg_edge_3-1
19535	lcg3+_gad_4					12728	lcg3+-8-2
						7847	lcg3+-8
						7419	Lcg_1

top 758 genes yielding the highest enrichment score for any single motif present in at least 500 promoters.

In addition, beyond these computationally derived motifs, the motifs of known human and mouse transcription factors in the form of position weight matrixes (PWM) from Jaspar [198], and select motifs from Transfac [199] and Onomine[129] databases, were also analyzed to examine the potential for known transcription factor binding pwm's to predict PRC2 binding sites (**Appendix 4.7**). On average, these known motifs performed approximately the same as the predicted motifs; the average

enrichment of best 10 motifs for these is 2.36 vs 2.31 for computationally derived motifs, with the highest of these from Jaspar being the transcription factor NFKB1 which was found to be present in 2012 promoters in the genome and has an enrichment of 2.55 against the top 758 genes (**Appendix 4.6e**). The highest for the other databases is LHX3 from Oncomine with an enrichment of 2.7 against 1407 promoters. All in all, 173 unique motifs were tested for enrichment (**Appendix Table 4.7**).

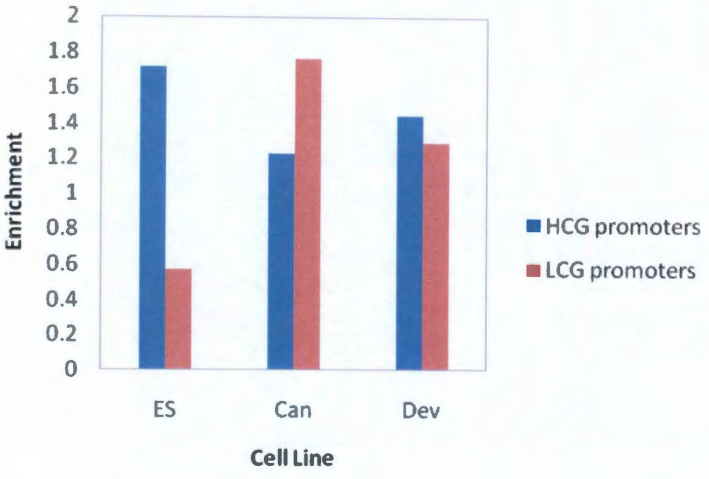


Fig. 3.5. Average enrichment of PRC2 marked genes in Embryonic Stem, Cancer, and Developed cells for high (blue) and low (red) CpG content.

Features Beyond the Motif Level

After initial analysis on the enrichment of the motifs, there could be seen a distinct pattern between the ESC group of experiments and the Cancer group of data. This first revealed itself in the previously seen phenomena in which PRC2 occupancy has been seen to closely match with the presence of CpG Islands [200]. This is bore

out in our analysis as well, with the class of ESC experiments beings enriched for CpG Islands by 1.41. Additionally, recent research has described a bifurcation in the definition of CpG Islands into so called Low CpG content (LcG) and High CpG content (HcG) Islands [201] and have shown that HcG Islands in particular are even more strongly correlated with CHIP results for PRC2 in ESC's [202]. This on aggregate is also true, with the enrichment on all ESC CHIP results being 1.72, an increase over all CpG Islands. What has not been previously observed is that while for promoters that have LcG Islands the a particularly low enrichment in ESC's, 0.57, the enrichment of this class in Cancer cells on aggregate remains very high at 1.77, and is even much higher than that of the HcG class at 1.23 (**Fig.3.5**). This indicates that while in ESC's HcG Islands are marked by H3K27me3 or PRC2 and LcG Islands are not, in Cancer, while HcG's remain somewhat marked, (due to their overall importance as regulatory genes), LcG Islands drastically increase their H3K27me3 or PRC2 binding, with an aggregate change in enrichment greater than 3x difference between ESC's and cancer cells.

This raised the interesting question of what could be special about LcG Island containing genes that would lead to them being so differentially modified. Gene Ontology analysis reveals that LcG genes are particularly enriched in the categories of stimulus response and extracellular processes [201], while HcG genes are enriched for the nuclear and transcription related processes [201]. This is also found when LcG's and HcG's specific to PRC2 recruitment are analyzed. This does perhaps indicate some mechanisms by which Cancer cells may be altering their extracellular response, but whether this is cause or affect is difficult to say, as it is known that

cancer produces a unique microenvironment to which the cells must adapt[203]. Consequently, to better understand how sequence could affect this, the motif enrichments were reanalyzed, this time to identify the motifs that showed the highest differential pattern between ESC regulated genes and Cancer cell regulated genes (**Table 3.13**). This shows that there are several motifs that behave in a similar manner as LcG with higher enrichment in Cancer cells then ESC's, while the bulk of the motifs show greater enrichment toward ESC's and thus could potentially indicate how genes might be recognized by different molecular programs. This observation that LcG Islands may be important in cancer is also seen looking at expression patterns. The database curated by Oncomine [129] has lists that they have generated from expression analysis across the Gene Expression Omnibus and have created their own series of top 10% overexpressed and underexpressed genes for given cancer or cell lines. Each of these lists yields approximately 500-2000 genes that are statistically up or down. A semi-random survey of 21 underexpressed cancer tissues and cell lines across Breast, Lung, Prostate, and Gastric cancers against 15 overexpressed cancers reveal a statistically significant increase of 1.5x (.82 into 1.26; p-val = $2x10^{-6}$) in the enrichment of LcG Islands from overexpressed to underexpressed, which makes since when we factor in that the increase of LCG's in Cancer CHIP data reflects the potential for increased repression from PRC2. Additionally, this shift goes from a significantly under-enriched state to a significantly enriched state, suggesting a real transition is occurring.

Motif analysis then shows that a group of motifs, which all have LcG enrichment and predict PRC2 occupancy (**Appendix Table 4.6f**) might be good

candidates for future experimental examination. Of particular interest among these, is the set of proteins which make up the Core Binding Factor Complex, RUNX2(osf2), PEBP, and AML2. These proteins have been shown to be oncogenes, which are critical in the development of the osteoblasts, but are also linked to the tumorgenesis of a number of different cancers [204-205], and RUNX3, another member of the complex (but not measured in motif analysis), additionally seems to be regulated by EZH2 creating the potential for feedback [206]. The binding sites for these proteins, particularly as taken as an intersection of different combinations of the three represent one of the few examples in which a factor has no significant enrichment against ES cell lines but which gains enrichment specific to Cancer cell lines (**Table 3.12**).

The reverse conditions were also explored to identify which factors might be the most responsible for the ES cell phenotype to the exclusion of cancer or differentiated cells. In this case, the number of CpG islands is seen to be the highest predictor with promoter regions with having two CpG islands being 2.05 times and having three or more CpG islands 2.25 times the average enrichment in ES over Cancer cells. Of the motifs that would seem to predict this pattern, the highest differential are found amongst those motifs which have the highest enrichment for HCG's, particularly NRF1, NFKB1, TFAP2A, CREB1 and p300, these being on average 1.6 times more enriched in ES cells than in Cancer cells.

Table 3.12. Features containing the highest differential enrichment between ES and Cancer cells

Highest differential for ES			**Highest differential for Cancer**		
Number of genes	Feature	es/can	Number of genes	Feature	es/can
190	UCSC_CpG-3+	2.25	3692	LCG promoters	0.32
1325	UCSC_CpG-2+	2.09	925	(up)C.A. vs. Norm.	0.67
1135	UCSC_CpG-2	2.05	1420	Osf2	0.68
64	znf354chighcon	1.79	1462	PEBP	0.71
2012	NFKB1-highcon	1.66	1938	758m3high	0.75
1466	motif6-highcon	1.63	1404	IRF (A)	0.76
6065	nrf1-con	1.62	1139	core-binding factor	0.78
6378	TFAP2A-highcon	1.60	1446	AML	0.78
4194	NFKB1-con	1.60	1957	(up)C. Adenoma vs. Norm.	0.79
3073	mat4-1cons	1.59	5147	gklfhigh	0.79
3918	p300highcon	1.58	618	(up)D.B.C. vs. Normal	0.80
5753	motif2-highcon	1.58	1404	Pit-1	0.80
3831	motif1-con	1.55	1957	L. - Top 10% O.E. (Wo.)	0.82
7611	10md-highcon	1.54	68	Brachyury	0.83
487	SMTTTTGT	1.54	1250	Nrf-1	0.84
3484	CREB1-highcon	1.54	1452	c-Rel	0.84
4514	nrf1-highcon	1.53	7847	lcg3+-8	0.84
1936	7gmd-highcon	1.53	1957	(up)G.A. vs. Norm.	0.85
5422	cpg_edge-3-2	1.53	7419	Lcg_1	0.85
2727	7gmd-con	1.52	136	SEF-1	0.85
8209	motif5-highcon	1.51	1492	NF-kappaB (p65)	0.87
1826	motif1-highcon	1.51	1481	Lmo2 complex	0.88
5600	10gmd3-highcon	1.51	1957	B.C. - Top 10% U.E. (Ad.)	0.89
2875	motif3-highcon	1.51	1375	B.C. - Top 10% U.E. (Gy. 2)	0.90
9095	pax4highcon	1.51	11702	lcg3+-6	0.91
9512	10md-con	1.50	1071	Lhx3(transf)	0.92
246	GCGNNANTTCC	1.50	8354	foxl1	0.92
5636	758m0highcon	1.49	12144	758m3	0.92
5030	mp_3-1	1.49	1381	NF-kappaB (D)	0.93
3660	motif6-con	1.49	1270	CHX10	0.95
4285	MZF1-4-highcon	1.49	12728	lcg3+-8-2	0.95

ES-Embryonic Stem cells; **C.A.**- Colon Adenocarcinoma; **L.**-Lymphoma; **G.A.**- Gastric Intestinal Type Adenocarcinoma; **B.C.**- Breast Cancer; **D.B.C.**- Ductal Breast Carcinoma; **C.**-Colon; **Norm.**-Normal; **Wo.**-Wooster; **Ad.**-Adai; **Gy.**- Gyorffy

Predictive Power and Evolutionary Comparison

While motif enrichment offers tantalizing clues about how local DNA sequence features could potentially alter the regional chromatin state by PRC2 mediated action, these motifs should also be able to convey some predictive power into either finding additional regions beyond what CHIP experiments can provide or offer evidence into which genes might be expected to be affected in a novel cell line. Drosophila PRE prediction has shown evidence that, using a set of 4-7 motifs, de novo PRE targets can be discovered which confer PRE behavior even while not standing out in CHIP [113]. Additionally, along these lines mouse PRE and motif prediction has shown that a similar approach may be of use in mammalian cell lines even while the specific mechanisms for PRC2 recruitment remain elusive. Unfortunately, in mammalian cell lines there are only two "true positives", one in mouse (KR), and one in human (HOXD11), so one has not enough data from which to truly train a predictive model and in affect the best you could do is try to predict CHIP targets. This carries with it a whole mess of assumptions and limitations that do not necessarily result in the most biologically relevant analysis. For this reason, instead of carrying out a fairly sophisticated model such as Bayesian Additive Regression Trees (BART) as was done by Yuan et al[207] or SVM, to classify the prediction potential of the motifs, instead only a simple linear summation model was chosen (see materials and methods). The motivation for this is that it does no good to create a best fit model against something to which you have less confidence in. However, we did wish to demonstrate that the motifs and genomic features studied here do possess predictive power on their own, that even with the simplest model is

comparable to what was demonstrated in the mouse system with the more advanced technique.

Using Area Under the Curve (AUC) analysis against an ROC plot one is able to measure the predictive power of the motifs. In this type of analysis a completely random prediction would yield an AUC score on .5, while a perfect prediction would yield a score of 1.0. For this type of motif prediction we are accepting that anything above 0.8 can be considered a "good" predictor, and as for comparison, Yuan et al using mouse data on a embryonic stem cell data set using primarily TRANSFAC motifs, was able to create a prediction model using BART, mentioned above, with an AUC score of 0.8266 with all 576 motifs factored in, and a score of 0.827 when the model is reduced to consider only 18 factors. Using our much simpler linear summation model utilizing 162 features yields a wide variety of predictive scores against the distribution of cell lines, with ES cell lines generally yielding "good" or near "good" prediction power while cancer cell lines yielding only poor predictions. The best prediction result is from a H9 ES cell line against suz12 antibody, which yielded an AUC score of 0.819. This number can then be increased slightly to 0.824 if non-motif factors such as CpG content and Oncomine expression data are used. Most of this prediction power is retained when paired down to only 14 factors (tlx1-nfic, rest, cpg_edge-3-2, Lcg_8, NFKB1-high, MZF1-4-con, TFAP2A-highcon, nrf1-high, 10gm2-con, 10md-con, motif2-highcon, motif5-con, pax4highcon, and 758m0highcon) (**Fig.3.6**) yields a score of 0.795, and if three more features (HCG promoters, UCSC_CpG-2+, and underexpressed-Colon Adenocarcinoma vs. Normal) are added the score becomes 0.807. This would suggest that these 17 features are the

most important determiners of SUZ12 localization in the promoter region of H9 Embryonic Stem Cells.

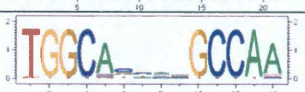
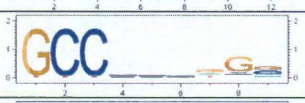
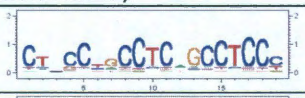
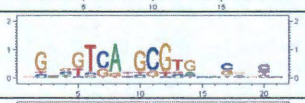
10md	
REST	
Tlx1-nfic	
LcG_8	
NFKB1	
MZF1-4	
TFAP2A	
nrf1	
10gm2	
motif2	
cpg_edge-3-2	TGCGCGTGCGCG (allowing for 2 substitutions)
motif5	
pax4	
758m0	

Fig. 3.6. Position Weight Matrix depictions of the 14 motifs that yield the highest predictive power toward Human ES cell prediction.

Additionally, it is of interest to see how this prediction compares to that of the mouse data. Yuan et al found that their model of mouse prediction when applied to a lower evolutionary organism, Drosophila, resulted in generally poor prediction against Drosophila CHIP experiments, AUC score equals 0.667 (**Fig.3.7**). As such, all of the motifs that went into the prediction of the human H9 cell line were mapped to the mouse genome (mm8) with the exact same parameters in CisGenome as that for the human genome. Applying this data using the same prediction parameters that were used above onto the mouse ESC CHIP used by Yuan et al, then yields a decreased score of 0.77 (**Fig.3.7**). This score, though lower, suggests that there are still significant similarities between mouse and human PRC2 at the motif level, even while there seems to also be some evolutionary divergence as well, given the suggested differences between human or mouse, and drosophila prediction. To put into perspective 0.77 is approximately the same score as will be generated applying this prediction criteria to one of the H9 H3K27me3 data sets (shown in position 6 of Table 3.2), which is the lowest scoring of all the ES CHIP sets with this prediction criteria.

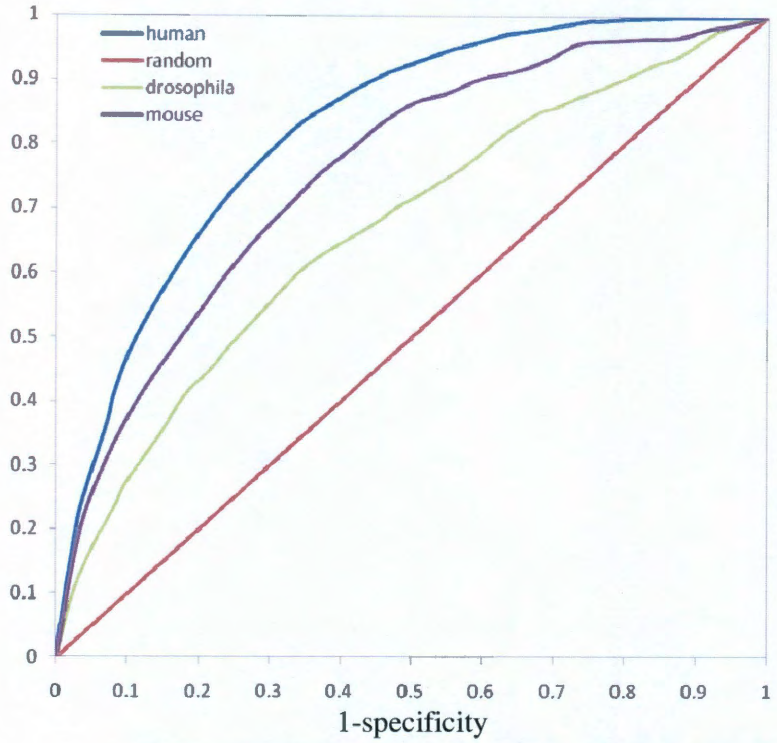


Fig. 3.7. ROC curve for prediction results using human motif enrichment against human (blue), mouse (purple), and drosophila (green) compared to random prediction (red).

3.4.2 Discussion

In order to better understand PRC2 binding site predetermination across the human genome, we have taken a broad overview of the existing landscape of CHIP-Chip and CHIP-seq experiments with the intention of identifying commonalities at the motif and DNA feature level within the promoters of the genes identified therein. While we acknowledge that there are likely other genomic regions beyond just the gene promoters that are critical to H3K27me3 regulation, this framework is the most convenient toward reconciling the experimental differences in the literature with the need to create a consistent structure from which to apply our computational methods.

This approach has been to combine computationally derived motifs using multiple motif discovery programs with the motifs of known transcription factors from publically available databases, along with CpG island content and expression data to identify those features which best describe the CHIP experiments across a spectrum of different cell types which can be loosely defined as Embryonic Stem Cell, Cancer Cells, and Differentiated Cells. This analysis has identified several features that would seem to indicate that the broad genomic programs that are regulated by PRC2 have different properties between ES cells and Cancer cells, most notably in the type of CpG islands that are being modified, with LcG's showing significant increases in enrichment among Cancer cells lines over ES cell lines, while HcG enrichment is decreased. Additionally, there are several motifs that yield ES specific, or Can specific enrichment (**Table 3.12**). Also by comparing the overlap between each CHIP experiment with every other experiment (**Table 3.13**), we find that ES cells as a group are on average 3.25 times more enriched for the genes of other ES experimentally identified genes than by Cancer experimentally identified genes and 2.6 times more enriched over those identified in Differentiated experiments. We also see here that while Differentiated experiments on average are actually more enriched in ES experiments than Differentiated experiments, Cancer and ES are both more enriched amongst themselves than as compared to any other group (**Table 3.14**).

Table 3.13. Comparison between each cell line studies with all others. Cell lines are grouped according to their class. The first section represents Embryonic Stem Cells, the next section represents Cancer Cells, and the third section Developed Cells. All values are presented as enrichments.

# genes in exp.	2169	3676	1040	1066	1121	2687	1435	5508	1351	1293	113	548	2118	4971	9375	11088	9423	2333	1867	2412	638	1385
Cell Line:	hES	H1	H9	H9	H9	H9	Ntera2	PC3	SW480	MCF7	G.A.	RL	RL	EP156T	EP156T	EPT1	EPT2	D.C.	Mac.	Mon.	TIG3	TIG3
Antibody used:	H3K27	H3K27	Suz12	EED	H3K27	H3K27	Suz12	H3K27	Suz12	Suz12	H3K27	Suz12	H3K27	H3K27	H3K27	H3K27	H3K27	H3K27	H3K27	H3K27	Suz12	H3K27
hES-H3K27	1.00	3.91	6.17	5.85	6.47	2.50	3.20	1.90	1.34	1.46	2.00	1.59	1.33	2.13	1.50	1.34	1.47	2.19	2.42	2.30	3.80	1.23
H1-H3K27		1.00	4.51	4.21	4.79	2.17	2.86	1.93	1.15	1.29	1.33	1.66	1.53	2.07	1.54	1.45	1.55	2.47	2.72	2.60	3.30	0.95
H9-Suz12			1.00	14.44	13.99	2.96	4.24	2.49	1.77	2.28	1.45	2.05	1.69	2.68	1.73	1.51	1.66	2.82	3.15	2.81	6.96	1.61
H9-EED				1.00	13.75	2.79	3.35	2.37	1.43	1.75	1.59	2.04	1.67	2.53	1.67	1.47	1.59	2.43	2.76	2.45	6.54	1.40
H9-H3K27					1.00	3.14	3.71	2.55	1.53	1.78	1.34	2.29	1.85	2.71	1.72	1.50	1.67	2.74	3.13	2.83	6.76	1.55
H9-H3K27						1.00	1.80	1.08	0.69	0.77	1.75	1.46	1.50	1.17	1.62	1.45	1.60	2.47	2.75	2.57	2.07	1.35
Ntera2-Suz12							1.00	1.68	4.02	4.69	0.39	1.57	1.46	1.83	1.30	1.24	1.28	2.15	2.34	2.25	3.16	1.11
PC3-H3K27								1.00	1.40	1.55	1.23	1.19	1.35	2.10	1.80	1.64	1.73	1.88	1.92	1.81	2.05	0.64
SW480-Suz12									1.00	5.75	1.67	1.47	1.64	1.49	1.07	0.98	1.08	1.23	1.35	1.08	3.36	1.18
MCF7-Suz12										1.00	0.44	1.23	1.87	1.59	1.13	0.98	1.09	1.57	1.65	1.41	1.39	0.65
G.A.-H3K27											1.00	1.03	0.89	1.17	1.29	1.19	1.10	1.53	1.31	1.33	0.59	0.41
RL-Suz12												1.00	5.14	1.44	1.04	0.96	1.08	1.53	1.71	1.50	2.19	2.02
RL-H3K27													1.00	1.65	1.24	1.10	1.26	1.65	1.75	1.47	1.89	2.18
EP156T-H3K27														1.00	2.13	1.70	1.87	1.93	2.00	1.83	2.39	0.85
EP156T-H3K27															1.00	1.75	1.82	1.59	1.58	1.55	1.38	1.22
EPT1-H3K27																1.00	1.72	1.48	1.48	1.47	1.20	1.08
EPT2-H3K27																	1.00	1.56	1.58	1.53	1.39	1.24
D.C.-H3K27																		1.00	8.47	7.20	2.22	1.65
Mac.-H3K27																			1.00	7.74	2.65	1.84
Mon.-H3K27																				1.00	2.24	1.58
TIG3-Suz12																					1.00	5.30
TIG3-H3K27																						1.00

G.A.-Gastric Adenocarcinoma; D.C.-Dendritic Cells; Mac.-Macorphages; Mon.-Monocytes

Table 3.14. Similarity of PRC2 marked genes between cells of different classes

Cell Type	ES	Cancer	Developed
ES	6.11	1.88	2.34
Cancer	1.88	1.98	1.49
Developed	2.34	1.49	2.28

In addition to the characterization of CHIP marked promoters, we also have shown that a simple linear addition model of motif, at this stage in our understanding of PRC2 function, is probably sufficient to create working lists of likely candidate genes from which to explore experimentally; as this model performs approximately as well as the more complex BART model proposed to analyze mouse ES CHIP data. Analysis of this mouse data also suggests that at the multiple motif level at least, there isn't significant sequence difference between the genes being regulated by PRC2, as the application of the motifs from the human model performs only as poorly in identifying mouse targets as the worst of the human predictions in a similar ES cell state. The above evidence, combined with the fact that drosophila predictions have proven very poor at picking mouse targets, would suggest that the localization of PRC2 components throughout the genome and across a large evolutionary distance perhaps highly malleable, but it would suggest common features are at play in the much more closely related mammalian species. This does not rule out that localization could be driven at the macro level, perhaps toward chromatin state, rather than potentially at the motif level, thereby allowing PRC2 to interact with multiple classes of transcription factors as indicated by seeming importance of high and low CpG content CpG Island presence regardless of sequence. Nor does this necessarily rule out the ncRNA theory of PRC2 recruitment, as in this case specificity is gained through complementation, and therefore

every binding site could have slightly different complementation characteristic which would not even necessarily be conserved in position across species. However, it does lend credence to idea that there are currently unidentified proteins, perhaps acting on the motifs identified above, which provide the basis for PRC2 recruitment and the resulting modification to gene expression mechanisms and that these proteins are probably conserved between humans and mice.

Consequently, the identification of the motifs and feature focused on here will allow for better search criteria in the experimental identification of Bona Fide mammalian PRE's, and provide further insight into Cancer Biology and how the regulation thereof represents fundamental shifts over what would be considered "normal" cells, and it is through these differences that better treatments can be found.

Chapter 4. Appendix

Appendix Table 4.1. Validation gene lists from Schwartz et al. 2006, Tolhuis et al. 2006, and Schuettengruber 2009. The validation gene lists from three experimental studies [75, 85-86]

Schwartz 2006			
bi	sob	Optix	Prat2
peb	odd	ana	CG32388
Vsx2	slp1	unpg	bin
Vsx1	slp2	inv	Doc3
CG4766	H15	en	CG5194
mab-21	mid	vg	Doc2
CG9650	sens-2	Psc	Doc1
ct	wg	Su(z)2	klu
oc	SoxN	Sox15	CG4328
Lim1	Osi21	Oaz	CG32105
lz	salm	kn	toe
btd	nub	fus	eyg
Sp1	ref2	grh	CG32102
disco-r	pdm2	CG7229	ara
disco	pburs	rib	caup
unc-4	noc	otp	mirr
OdsH	esg	CG9235	bru-3
slgA	wor	Rx	Sox21b
zfh2	CG5888	dve	D
ey	dac	retn	HGTX
Gsc	CG15167	bs	CG8765
chinmo	ham	Dll	knrl
CG15357	CG10570	gsb-n	kni
CG31670	CG17325	gsb	croc
CG3597	tup	CG16778	AP-2
CG3609	ssp3	Kr	CG14658
CG9886	dia	trh	CG14659
dpp	cad	CG13891	opa
toc	ap	bab1	lab
CG8853	CG1701	bab2	Edg84A
drm	so	vvl	pb
Actn3	lbl	CG11920	zfh1
Hmx	lbe	fd96Ca	Ptx1
CG18139	C15	fd96Cb	Sox100B
sr	slou	fkh	abd-A
tin	CG15498	Dr	Abd-B
bap	hh	dmrt99B	CG14909
zen2	Antp	hth	svp
zen	grn	KP78b	sim
Dfd	Poxm	KP78a	E5
Scr	Or85a	pros	ems
ftz	CG7443	CG3942	GATAe

Tolhuis 2006				
Pph13	Ucp4B	nimB4	dia	CG1701
Gsc	eya	nimB5	cad	so
ds	nrv2	He	Pomp	CG11145
lea	CG17376	nimC1	CG31612	CG11196
chinmo	CG17377	Cyp28a5	tsh	Hey
CG15357	CG11236	ppk	CG11629	CG11191
CG31670	CG17375	elB	CG1421	Odc1
CG10908	sens-2	pburs	CG1428	Odc2
CG31668	wg	Cpr35B	CG2528	CG14762
CG33124	Wnt6	CG15283	CG31693	Optix
CG15385	raw	noc	tio	CG12769
dpp	SoxN	CG4218	CG31601	CG8635
lilli	CG4382	CG3473	ap	ptc
CG8853	gcm2	mol	CG11163	CG13743
drm	Samuel	CG15269	Or42a	CG8197
sob	CG18666	esg	Tsp42A	ana
odd	ab	CG5888	Or42b	CG8083
CG34340	CG14926	Idgf1	EcR	CG11778
slp1	salr	Idgf2	jing	unpg
slp2	salm	Idgf3	CG15233	CG8027
H15	Pde1c	dac	CG15234	wun
CG31647	bru-2	CG4580	CG9422	wun2
mid	nub	CLIP-190	Tdc1	Mef2
CG13999	ref2	CG15167	CG15909	eve
CG13998	pdm2	CG10570	Tdc2	Pka-R2
Vm26Ab	CG15485	CG17325	Tsp42Eg	psq
Vm26Ac	CG31856	tup	Tsp42Eh	CG11883
Vm26Aa	B4	ssp3	Tsp42Ei	sprt
psd	nimA	tj	CG12842	CG7777
CG13992	nimB1	CG17571	Tsp42Ej	qvr
Ucp4C	nimB2	CG17570	Tsp42Ek	E(Pc)
CG32812	fus	arg	inv	CG17580
DAAM	CG8207	elav	CG30034	vg
br	Zasp52	CG4293	en	Sans
CG3600	Poxn	Appl	Roc2	CG30487
csw	sli	vnd	CG30035	Mdr49
ph-d	ci	CG13366	CG8234	CG3884
ph-p	ey	sdk	Ir48b	CG13321
CG3835	bt	CG13362	CG8550	Psc
Oaz	CG11231	CG13361	CG12370	Su(z)2
kn	Sox102F	CG5254	CG8776	CG13323
CG10200	CG11152	fz3	CG12442	CG13324

ss	Ubx	CG31275	CG17631	CG10202	toy	tw	CG12374	Prosap
Fas1	CG31498	Glut3	pnr	unc-5	sv	A3-3	sca	CG42287
				CG30069	CG8394	Sox15	CG17388	CG17389

Schuettengruber 2009

abd-A	CG14974	CG7443	en	inaD	pnr	SoxN
Abd-B	CG15167	CG8765	esg	ind	pnt	Sp1
al	CG15269	tey	Ets65A	inv	Poxn	sr
Antp	nub	CG8853	eve	klu	Prat2	srp
ap	Hr51	CG9235	exex	kni	prd	ss
AP-2	Oaz	CG9650	ey	knrl	pros	Su(z)2
ara	CG17631	cnc	eya	Kr	Psc	sv
Art9	ssp3	CR32730	eyg	lab	Ptp61F	svp
Awh	CG2014	croc	fd3F	laf	Ptx1	tin
B-H1	CG2052	ct	fd59A	lbe	repo	tio
B-H2	CG2543	Cyp313a3	fd96Ca	lbl	retn	tko
bi	Sytbeta	D	fd96Cb	Lim1	Rfx	CG16778
bin	CG31275	dac	fkh	lz	Rh5	tll
bs	CG31386	Dfd	fzo	mab-21	rpr	toe
bsh	sens-2	disco	GATAe	mid	run	toy
btd	CG31670	disco-r	gl	mir-276b	Rx	trh
bxd	CG32105	Dll	grh	mir-iab-4-3p	salm	tsh
cad	CG32111	dmrt99B	grn	mirr	salr	tup
caup	CG32532	Doc1	gsb	noc	Scr	twi
CG10349	CG32713	Doc2	gsb-n	oc	sens	Ubx
CG10570	CG33325	Doc3	Gsc	odd	sim	unc-4
Fie	CG33797	dpp	gt	OdsH	Six4	unpg
CG11023	CG33798	Dr	H15	opa	slou	vg
CG11629	Vsx2	drm	ham	Optix	slp1	vnd
CG12684	CG3835	dsx	hbn	Or67d	slp2	vvl
CG12685	Vsx1	dve	hdc	os	so	wg
CG13321	CG4766	dys	HGTX	otp	sob	Wnt6
CG13891	CG5718	E5	hh	pb	Sox100B	zfh1
CG13996	Hmx	Edg84A	hkb	pdm2	Sox102F	zfh2
CG14574	CG5888	eg	hth	peb	Sox15	
CG14659	CG6023	ems	iab-4	ph-p	Sox21b	

Appendix Table 4.2. List of top 243 PcG target genes predicted by *EpiPredictor-Basic*

Gene Name					
Rbp6	pum	CG6220	Pdp1	Bx	grh
hth	salm	tutl	sm	CG17839	CG32547
Antp	Ptp99A	Sdc	SKIP	CG10793	pdm2
CG34354	unc-13-4A	disco	CG10908	CG15465	rut
Abd-B	H15	shot	pnr	CG32720	shep
beat-VI	CG2750	htt	GluClalpha	Obp47a	CG5778
bi	CG12540	CG33988	Cyp12e1	opa	beat-Vb
br	Eip74EF	CG7722	os	CG11085	CG14532
kek5	Ubx	tna	mir-289	B-H1	cav
CG34353	CG30350	chinmo	OdsH	CG18208	CG14298
mbl	sif	CG18371	CG12637	ftz-f1	mir-280
nAcRalpha-7E	Tl	dpr13	mab-21	CG32494	cbt
bru-3	CG15464	noc	CG1631	ImpL3	sog
luna	rg	tlk	stops	ham	toe
eag	disco-r	sdt	brk	HGTX	app
abd-A	CG34362	Sema-1a	CG32698	CG12605	Ptr
tou	Cnx99A	CG14826	ems	jing	cv-2
olf413	nAcRalpha-96Aa	pros	CG42342	CG6734	CG2865
CG42611	CG12484	zfh1	rib	Rya-r44F	E5
ct	corto	Appl	CG30115	eve	CG9134
tomosyn	CG15025	CG9571	px	CG5142	CG34393
fas	Imp	tok	PhKgamma	Fas3	pdfr
CG3600	Dr	eya	CG2444	CG6490	osa
Vsx2	SoxN	dx	aPKC	CG30089	
lilli	CG33691	wg	Dll	CG5075	
fz2	aret	gukh	vvl	CG6424	
ds	Idgf4	nudE	CG6123	CG14985	
CG14621	Dl	tsh	CG32204	chic	
Prat2	SK	hdc	Poxn	CG15198	
Mur2B	Optix	lola	nkd	sens-2	
CG9650	B-H2	Ten-m	CG31714	CadN	
shn	CG12877	lea	NFAT	ETHR	
caps	nAcRalpha-30D	CG42339	cpo	CG11997	
CG32635	Vsx1	ps	Eip93F	CG32111	
CG42340	unc-4	CG10349	Obp85a	sano	
CG17388	Dys	CG32105	scyl	CG42265	
ss	CG31386	Sp1	lds	wupA	
en	grn	Gld2	Lar	Caki	
r-cup	CG17230	Rh5	Eip63E	CG9308	
dpr6	CG32541	D2R	ru	Ca-alpha1T	
CG9059	ush	scrib	CG32193	CG31498	
side	cad	Ets98B	Cda4	CG11486	
slou	dpr8	Pde9	Glu-RIB	CG30377	
Ten-a	RunxB	CG31128	Ir41a	CG17018	

Appendix Table 4.3. List of top 322 PcG target genes predicted by an advanced version of *EpiPredictor* with comparative genomics integrated (*EpiPredictor-CG*)

Gene Name				
Rbp6	fas	SoxN	GluClalpha	chic
hth	Vsx2	Idgf4	osa	wg
CG34354	tutl	CG17839	mab-21	shot
mbl	Cyp12e1	Vsx1	inv	Dl
abd-A	CG42340	ems	CG17208	dpr8
CG42611	bru-3	CG32264	CG34347	orb
br	aret	Lar	InR	Dgk
ds	shn	tna	pros	gukh
luna	lea	CG31145	pnr	rdx
CG3600	ush	Ca-alpha1T	pum	fig
beat-VI	CG13235	Ptx1	Ten-m	upd3
H15	ss	CG32635	jing	tomosyn
hdc	PQBP-1	sens-2	CG12551	CG15395
olf413	noc	Ir94f	Con	CG5075
en	htt	CG18478	spri	aPKC
Ubx	salm	apt	ctp	cbt
side	ftz-f1	CG9650	pdm3	CG31714
Abd-B	Antp	trh	Pdp1	Ser
CG34362	zfh1	tlk	Ten-a	Dfd
cpo	Sdc	elB	CanA-14F	CG34360
sdt	eag	CG10814	Tbh	CG9059
Dr	grn	unc-4	sog	CG32547
fz2	CadN	Mmp2	rut	SKIP
Sp1	lilli	Dll	CG32541	CG32398
CG32698	CG1499	Appl	B-H2	CG30115
CG31386	Eip63E	CG32613	Mrtf	drl
SK	CG15464	disco	CG42339	CG13830
px	tup	Ets98B	opa	unc-13-4A
Glu-RIB	mub	Prat2	DopR2	mtt
Sema-1a	D2R	RunxB	CG42594	beat-IIb
CG12484	CG34353	CG14532	CG31827	Tl
dpr6	Gld2	Obp85a	nkd	ps
mir-289	cad	Eip74EF	CG30083	Rh5
HGTX	CG10349	dpr13	app	bi
CG6220	CG40006	fz4	pnt	CG9571
slou	ct	Obp47a	toe	CG31498
nAcRalpha-30D	shep	CG18208	vvl	Fas3
CG7722	sif	DopR	nAcRalpha-7E	sm
chinmo	otp	caps	tsh	ara
CG32494	Caki	stan	Eip93F	CG10126
beat-IIIc	dpp	corto	beat-Vb	eya
CG12187	rib	dac	CG32521	brk
psq	caup	run	CG17230	Poxn
CG9817	tok	CG12540	nub	CG32062

Pde9
beat-VII
ham
trio
Gsc
r-cup
Dys
salr
Cpr76Bd
dx
sca
CG2014
Glut1
Ptr
brat
Optix
Lim1
CG10908
CG10793
CG7470
dlg1
CG32111
ETHR
CG33298
nAcRalpha-96Aa
CG10384
CG9134
NFAT
CG14298
cenG1A
CG33988
beat-IIa
lbl
CG31337
KrT95D
CG12835
CG18371
bab1
CG18482
stops
CG13972
prod
ade5
CG15630
Rya-r44F
mAcR-60C
Mvl
CG14509
CG3650
CG14521
slo
Vha16
cic
Cda4
os
CG6123
dpr15
Hex-A
CG30350
CG10830
Wnt10
CG14909
CG4669
CG31235
Takl1
CG13872
CG17048
CG4372
CG15465
wupA
CG32105
CG12538
oc
CG4476
Hip14
CG16716
pdm2
fj
CG4168
CG8888
CG11085
CG1986
Mur2B
elk
eyg
Vha16-3
CG9308
CG15233
CG11298
Teh1
brp
scyl
for
CG14826
CG11486
abba
ImpL3
vg
CG6175
ru
kek1
CG5142

Appendix Table 4.4. Genomic coordinates and genes, along with the primers, used for qPCR.

A. 15 selected genomic regions from prediction

Closest Gene	Genomic Coordinates	Prediction Rank	Primers (top-foreward)
Antp	3R:2826281..2826940	18	GGCATCCAAACATCCACTTG TTATGAGTGTGCGTCTGTGG
disco	X:16110941..16111700	9	GTTTCGTTGGGTTGACACATG CATTGCCATTTCACTCTCGTTG
eve	2R:5865941..5866860	15	CGGCATAATATTAAGACTTCA TCCCACTATATATTTGTATGTATG
unc-4	X:17662101..17662800	10	GGCTGATCGAAATTGAAACGG AGCGAGGAAACCCAGAAAAG
ubx	3R:12589661..12590180	25	CTGTATCTCGCTCTTACGCAC CAAAACACGAATACAAGCCCG
salm	2L:11445481..11446280	2	CACTATCACTCAGCCAACCC ATCCCGAGGCAAAAGTAGAAG
noc	2L:14490381..14491000	8	TTACAGGAAGCCAAATCGGAG GTCCAATCACAATCGCATGC
pnr	3R:11851381..11851900	23	TCTCTTGCTCTTTCGCTCAC GTTTTCCATACGCACTCACAC
grn	3R:3977421..3977780	123	CACAGCTCGAAATGACAAACG TTCGCTGTCTCTTTCACTGG
CG13896	3L:723581..723960	119	TCTCTGTCCCTCAAGCTATGG CTGTAGAAGTCCAGCTGTTCAG
abd-A	3R:12637441..12638040	64	CCCATAAATCACGACTCCCAG TCGCTCAGATCCACATTCAC
bab2	3L:1127161..1127580	57	TTGGCCTCGACTGTTGATG GGCAAAGAAAAGTTGGGTGG
CG12484	2R:16326581..16327020	132	CCACTAGCCCATAACAGTAACAG TTCGGACTGGTCGTTTATGG
zfh1	3R:26589841..26590200	77	GGAAAATGGCTGGGAAAATGG TGGAAAATGTGAGAGCAGGAG
CG4774	3R:21514901..21515540	28	CCTAGTGAAGATGGCTAACGTC GTATAGAGATGGCCGCTAATGG

B. Control Genes used in qPCR

Gene	Control Type	Primers (top-foreward)
Hsp22	Negative	TCTGCGTATGGAAACTGACC TTCTTTAGCGAACTCCGTGG
Hsp26	Negative	AATAGTGGGAGATTGCTGGC CCTTTCCCAATAAATGCCATGAG
Pc	Negative	CTATTCCATTGTCCTGTTTGCG ACGTCAAAACTATGAGAGGCG
bxd	Positive	CAAAACACGAATACAAGCCCG CTGTATCTCGCTCTTACGCAC
iab2	Positive	TCGCTCAGATCCACATTCAC CCCATAAATCACGACTCCCAG
en_DM	Positive	GGAATGGGTAAGAGGAAGATGG AACTGGAAACTGGAACGGAG
hth	Ringrose	ACCGCCATAATCTTGACAGAC GCGTTGCCATAAAACACTTAGG
unc-4	Ringrose	GGCTGATCGAAATTGAAACGG AGCGAGGAAACCCAGAAAAG
ldgf4	Ringrose	AAGGCGAGAGGGAGATAGAG ACATTTTCACCAGGACAGGG
cato	Ringrose	CAAGTTTGTGTAAATGGCCCG AAATTATGGGCGACAGAGGG

Appendix Table 4.5. Genes present in at least 50% of CHIP experiments

CBX2	SCNN1B	SST	TBX22	DPY19L2	DPYSL5	CTNND2
CGNL1	C18orf1	TMEM2	TMEM65	EMX1	FRMPD1	DHH
GABRQ	EPN3	TMEM26	ABO	HOXD12	HOXC6	DLX1
FHL1	CPAMD8	EDN3	ALK	HS3ST4	HOXD1	EOMES
PANX2	NPR1	AVPR1A	AVP	IGFBP1	HOXD11	FBP1
ADRA2B	PDGFB	C1QTNF4	BMPER	NKX2-5	KCNK17	FOXD2
CACNG7	RHCG	CGREF1	C1QL3	NPY5R	NGB	GABRA4
CNFN	ZNF232	FADS6	C1QTNF2	SLC17A7	NIN	GDF6
CUTL2	C14orf162	G0S2	CDKN2A	SMPD3	NKX6-2	GDNF
CYP19A1	CBLN2	GPM6B	CRH	TBC1D1	NTRK1	GHR
DPCR1	CHST2	GRIN2C	CSPG5	TCF15	OTOP1	GPR120
FAM70A	DACT1	LHX1	DAPK2	THSD3	POU3F1	GPR88
GPRC5A	DIRAS2	LHX9	DDX25	HLA-G	POU3F2	GRIK3
HAPLN4	GFRA3	NDRG2	EPO	CNNM2	SCUBE3	GSCL
JPH3	GIPC3	PREX1	FHL2	CNTNAP5	SOX1	HAND2
LGR6	HOP	RRAD	GPC3	DFNA5	SYT10	HBA1
NTF5	IGSF11	SEZ6	HS3ST1	EPS8	TCF21	HHIP
PRDM14	KAL1	SHD	LYPD1	GPR126	TFAP2D	HTR1A
PTPRO	KCNIP1	SLC6A4	MT1G	HOXA6	TMEM46	ICAM5
RARRES2	KLK1	SSTR4	MYO10	KCNJ10	CD38	INA
RDH8	LGICZ1	FGF11	NPTXR	MLPH	EGR2	ISL1
RPL38	LPPR4	RASD1	NR3C2	RUNX2	GFI1	ISL2
SRPX	LTF	SEZ6L	NRXN1	SLIT3	C2orf32	ITGA4
TRPC4	NTSR1	SLC13A3	PCDH10	TLE2	EPB41L4A	KCNK12
TUB	P2RX2	WNT5A	RNF157	FOXD4L1	FAM84A	KCNQ3
BMP7	PCDH7	CDYL	SATB2	MLNR	NFIX	KCNV1
CACNG4	PLCXD3	F13A1	SLCO4A1	OVOL1	ECEL1	KY
CACNG8	PVALB	GADD45G	SNTB1	RIPK3	LRRTM1	LHX4
CHAT	RASGEF1A	GRM2	TNFRSF13C	CACNA1E	PTHLH	MSX1
HCK	RDHE2	NKPD1	IRAK3	CBR3	BARHL1	MYF6
HOXC10	RPRML	OSBP2	NTN4	CD44	BCL2	NPR3
OPCML	RSPO3	RPS6KA2	ADAMTSL3	CLEC4G	CACNA1B	NTRK2
PTCHD1	SEMA5B	SMOX	CA7	CLIC5	CACNA1D	OLFML2B
RAI2	SNCB	SOX9	CART1	COL12A1	CNNM1	OSR1
GPR101	CITED1	HOXC4	MAP6	NR2E1	PMP22	SIX6
GJB2	C20orf103	COLEC12	KIRREL3	SSTR2	ZBTB16	ROBO3
HOXB6	HOXC8	NEFH	PLXNC1	POMC	TRPC5	DUOX2
NELL1	NOL4	PGR	PHOX2B	PTGDR	SLC32A1	VDR
SLC30A3	SLCO2A1	SLCO5A1	TBX5	TMEFF2	TRIM36	TSLP

OTX1	LRRK1	CACNA2D2	ACTL6B	NOG	CACNA1G	TBR1
PDGFRA	RPP25	COL9A1	FAM80A	NRN1	CD34	TCF2
PENK	ADAM23	CYGB	FGF20	NXPH4	CD8A	TIP39
PRDM12	APCDD1	FAM46C	GPR158	PCDH17	CENTA2	TLX2
PRKCE	ATP8A2	FOXC2	GPR26	PRDM13	CIDEA	WNT6
RBP4	BDKRB2	H1F0	IRF4	RAB40B	CSMD3	ZADH2
SFRP5	BMX	KCNS3	KCNJ5	RFXDC1	CYP26B1	ZIC4
SIX2	C14orf37	LIMS2	KCNK3	SEMA6D	DCC	CBX8
SLC10A4	FGF17	LIPG	MAPK8IP2	ST8SIA1	DKFZP564O0823	CH25H
SLC1A4	GALNTL1	NFATC1	RAB9B	TFAP2B	DOK6	COL24A1
SLC30A4	NMNAT2	PODXL2	RASGRP1	VGLL2	EN2	FBN2
SLC9A2	PAQR5	PTPN5	SORCS2	ZIC1	FAM5B	KLF4
SORCS3	PPP1R16B	RAB6B	EYA2	ADAMTS5	FGF3	MLLT3
SPOCK3	SLC5A8	SLC26A10	AQP5	ALOX5	FOXF1	PLXNA2
TFAP2E	SLITRK2	ST6GALNAC2	GDA	MESP1	FOXJ1	RASSF5
TLX1	CDH22	STAC2	DPF3	MT1A	GATA4	RGS10
TMOD2	CDH4	TSC22D3	EGFL6	NAGS	GIMAP5	STK32B
TRADD	GRIN2B	ACCN1	EPHA10	NFIA	GRID1	ZNF503
TRH	NRXN3	ABCC3	EVX1	PITX3	GUCY2D	PTGFR
WNT2	PRPH	CDH13	FOXG1B	PPP1R14C	HPCAL4	SORCS1
HEY1	SOX3	ERG	FOXL1	SFRP4	HPSE2	TCEA3
NAP1L2	FBXL16	GAD1	GABRA2	TRIM7	IRX4	SOX5
CCND2	PHOSPHO1	HOXA10	GSH1	ZNF365	KCNA3	ACCN4
C15orf27	ABCG1	HTR2C	HLX1	CHN2	KCNK2	ATP1B2
DUSP9	DPP4	KCNJ9	HLXB9	EIF4E3	MAPT	C20orf39
FGF13	EPB49	NEFL	HOXC11	PHYHIPL	NEUROD1	EPB41L1
NPAS3	FAM19A2	NOS1	IGF2	SLC35D3	NEUROG1	FAM81A
PNPLA5	PLXDC1	RNF128	KCNC4	LOC400120	NKX2-2	GAB3
RORA	AATF	TRHDE	KCNIP4	CXCL16	OCA2	GAL
MSI1	ADAM11	WNT3	KCNK4	HES7	PAX6	GNAS
ABCC4	ADAMTS17	FOXA1	MAL	TBX2	SIX1	KIRREL2
ADRBK2	ADAMTS8	INHBB	MSX2	ADCY8	SLC6A3	RGMA
C13orf18	ADCY1	LAMB1	MT1H	ADRB3	SPAG6	SHANK1
CACNG3	BRUNOL4	MTSS1	NGFR	ALX4	STXBP6	SLC7A10
ASCL2	DLX4	DMRT1	HS3ST3B1	LHX6	RTN4RL2	SHH
MAFB	MYOD1	NRG2	OTOP3	PAX3	PROK2	SIX3
GPR12	KCNA1	LHX5	WNT10B	NPAS1	ADCY4	ATOH1
CNTFR	EFNA1	ITPKA	SIDT1	ALX3	CDH7	COL25A1
ZFYVE28	RAB37	HBA2	RFX4	RGS6	ALOX15	VAX2

CCKBR	OLIG1	LYSMD2	CRTAC1	RYR3	KCNK10	FZD1
KCNN1	XYLT1	PRKCH	DGKG	SLC30A2	DGKZ	HAP1
SPTB	C1QL1	TNFRSF11A	DGKI	SLC6A1	ARHGAP6	LPL
KCNG1	FBXO3	GALR2	DMRT2	SLIT1	BCAN	METRNL
MN1	GRIN1	HS3ST2	DSC3	SLITRK3	BDNF	NRIP3
ALDH1A2	HOXA7	KLK4	FAM43B	SOX14	CHST1	PAX5
C1QL2	LMOD1	MDS1	FEV	SOX7	CHST8	PDE8B
CACNA2D3	MDGA1	PAX9	FGF9	SSTR1	FGF14	PSD2
CNIH3	MT1B	RBP7	FOXA2	TBX21	IL17RB	REPS2
COL23A1	NEURL	ADRA2A	FOXL2	WNT3A	SNIP	ST8SIA5
CRHR2	PTPRN2	BMP6	FZD10	ABTB2	DLK1	ABCC8
GREM1	RIMS4	COCH	GATA2	EPHB3	SPON1	ADCYAP1
GRM1	SPOCK2	COL27A1	HOXD13	MSC	PYY	BHLHB3
GRP	SYT6	CRHBP	HS6ST3	CHRD	ACCN2	CALCA
H2AFY2	KCNA5	DCHS2	KCND3	CLEC14A	KCNJ4	CHRDL2
NDRG4	CRMP1	FERD3L	LBX1	EFNA3	NCAM1	COL2A1
NRXN2	CRYBA2	FOXQ1	NEUROD2	ERBB4	NPHS2	CRHR1
NXPH2	FLJ33790	HSPA6	NEUROG2	FLRT2	NRCAM	DLL4
PAPLN	HLF	IRX1	NKX2-3	GALNTL4	DSCAML1	DMRT3
PPP1R1B	HOXB13	NKX3-1	NKX2-8	RGS20	ELAVL3	ELMOD1
SLC24A3	HOXB7	NRG1	NKX6-1	SNFT	HOXC12	GATA6
ZCCHC12	HOXB8	PTGER3	NPTX1	WNT7A	IL1RAPL2	LGR5
FOXB1	HOXD3	SLC40A1	ONECUT1	C1orf92	SIM1	NAV2
GPR50	HOXD9	SPON2	OTP	LTK	SLC6A2	OLIG2
HOXC9	KCNJ6	IRX3	OTX2	OPRD1	SLC6A20	ONECUT2
NTRK3	LGI3	ADAMTS15	PAX2	PODN	SLC8A3	OTOP2
PAX1	LHFPL3	ARHGAP20	PCDH8	SHOX2	TLX3	PHOX2A
PCSK2	MEOX2	CBLN4	PITX2	UCP1	DUOX1	POU4F1
SLC18A3	OLIG3	IRX5	POU4F2	SYT3	ZMYND15	POU4F3
CBX4	SV2B	ADRA1A	PRAC	C21orf29	ARNTL	SCN4B
PDE1B	TBX1	BARX2	PTF1A	ARX	PHLDB1	SIM2
PRKCB1	TBX20	BNC1	PTGER2	FXYD6	PRICKLE1	ST8SIA2
CLIC6	CCNA1	CA10	PTPRT	LONRF3	ATF3	WNT1
LAMA3	KCNQ1	CDK5R2	RAX	RPH3A	CACNB4	C21orf63
SLC1A2	WT1	ASCL1	COMP	CRLF1	HRK	PAX7
SLC24A4	VSX1	INSM1	WNT11	LRP2	KCNAB1	INSM2
BARHL2	COL9A2	CXCL14	CYP26A1	DIO3	GPC5	GRIK1
DACH1	EGR4	GHSR	GRIN3A	GSC	MCOLN3	NTNG2
ADAMTS18	MEGF11	SLC27A2	USH1G	ABCG4	CALCR	CAMK2B
SLITRK1	STMN2	RASGRF1	FLI1	HES2	KCNH3	KCNK13
HOXB2	HOXB3	NR2F2	PAX8	PDE4DIP	ASTN2	ATOH8
CDH23	EN1	EPHA5	EPHB1	FRMD3	GBX2	GRIA2

TBX3
TRIM9
UCN
CACNA1A
CACNG2
HOXB9
MYH11
TMEM16B
C20orf46
ATP1A3
CA4
CYP2A13
GJB6
PHACTR3
TBX4
RIPK4
WNT5B
ARHGEF7
HOXC13
ITGA11
UPB1
EMX2
HOXB1
RXRG
CYP24A1
CDX2
DLX3
TAL1
INSRR
KCNMA1
KL
LHX2
LMX1B
NEUROG3
NR4A3
SCTR
SGPP2
SLC35F3
SLIT2
ADRB1
BHLHB5
GUCY1A3

Appendix Table 4.6. All enrichments within an Enrichment Category that meet a threshold of 1.5. Categories are a) Embryonic Stem Cells, b) Cancer Cells, c)Differentiated Cells, d) The 758 genes that are present in at least half of all the CHIP experiments, e) Low CpG content CpG islands, f) High CpG content CpG islands

A. Embryonic Stem Cells

# of promoters		ES ave.						
758	top758	11.79	2083	motif4-highcon	1.80	3961	cpg_edge-1-1	1.62
190	UCSC_CpG-3+	4.19	232	znf354chigh	1.80	8530	nrf1-high	1.61
1325	UCSC_CpG-2+	2.73	2583	$TATA_01	1.79	7566	MZF1-4-con	1.61
1135	UCSC_CpG-2	2.49	1957	Rectal Adenoma vs. Normal	1.78	5422	cpg_edge-3-2	1.60
64	znf354chighcon	2.37	3967	$SP1_Q6	1.77	2632	deltaef1highcon	1.60
1407	$LHX3_01	2.23	7829	10gm4highcon	1.76	3203	LHX3_cons-5	1.59
1855	NRSF	2.22	860	Lung Adenocarcinoma vs. Normal	1.76	2700	LHX3-con	1.58
2012	NFKB1-highcon	2.16	6217	10gm2-con	1.76	203	mat_0cons	1.57
1936	7gmd-highcon	2.13	6199	10gm3-con	1.76	978	Invasive Ductal Breast Carcinoma vs. Normal	1.57
487	SMTTTTGT	2.10	3918	p300highcon	1.74	618	Ductal Breast Carcinoma vs. Normal	1.57
2091	17md-con	2.09	3660	motif6-con	1.74	1414	14md-con	1.56
1826	motif1-highcon	2.08	1094	Small Cell Lung Carcinoma vs. Normal	1.74	246	GCGNNANTTCC	1.55
1795	17md-highcon	2.07	8209	motif5-highcon	1.74	671	lcg3+-5	1.55
2727	7gmd-con	2.04	4514	nrf1-highcon	1.73	260	yy1-high	1.55
4194	NFKB1-con	2.02	165	NRSF	1.73	11029	Lcg_8	1.55
7476	cpg-1	2.00	9961	HCG promoters	1.72	8695	17ggm4-con	1.54
6378	TFAP2A-highcon	1.98	138	yy1-highcon	1.71	1262	Squamous Cell Lung Carcinoma vs. Normal	1.54
1466	motif6-highcon	1.97	1809	sef-1(cisgenome)	1.69	6059	motif1-high	1.53
5636	758m0highcon	1.93	9095	pax4highcon	1.69	1368	SOX9-highcon	1.53
1317	$AP1_C	1.93	5030	mp_3-1	1.69	2230	exsr1-fli1	1.53
3831	motif1-con	1.91	2389	tp53	1.69	925	Colon Adenocarcinoma vs. Normal	1.52
5753	motif2-highcon	1.90	6519	NFKB1-high	1.69	3413	ap1_c-highcon	1.52
3073	mat4-1cons	1.90	4167	lcg3+_gad_3	1.69	11175	nrf1	1.52
1957	Lymphoma - Top 10% Under-	1.88	1530	Prostate Carcinoma vs. Normal	1.68	6575	tlx1-nfic	1.52

	expressed (Wooster CellLine)							
4285	MZF1-4-highcon	1.87	5802	17ggm4-highcon	1.68	7780	HMX3	1.52
228	$NFKB_Q6_01	1.86	8512	motif2-con	1.68	4200	mp_7-0	1.51
2875	motif3-highcon	1.86	9512	10md-con	1.68	1060	HEN1 (A)	1.51
5600	10gmd3-highcon	1.85	8711	sp1-highcon	1.68	10963	motif2-high	1.50
1957	Colorectal Cancer - Top 10% Under-expressed (Wooster CellLine)	1.85	1409	Egr-1	1.68	11373	motif5-con	1.50
858	Prostate Carcinoma vs. Normal	1.85	451	$IRF_Q6	1.68	2384	lcg3+_gad_8	1.50
5032	rest	1.84	1297	LHX3-highcon	1.66	1364	Genes with at least one ERa binding site within 20kb of transcriptional start site	1.50
1572	$NFAT_Q4_01	1.84	3484	CREB1-highcon	1.65			
6330	7gmd2-highcon	1.82	6370	motif3-con	1.65			
6065	nrf1-con	1.81	1777	Prostate Adenocarcinoma vs. Normal	1.65			
4325	10gm3-highcon	1.81	4881	motif4-con	1.64			
4326	10gm2-highcon	1.81	2358	LHX3_cons-9	1.63			
7611	10md-highcon	1.81	9174	10gmd3-con	1.62			

B. Cancer Cells

# of promoters		Can average
758	top758	4.25
190	UCSC_CpG-3+	1.86
3692	LCG promoters	1.77
858	Prostate Carcinoma vs. Normal	1.76
136	SEF-1	1.72
1407	$LHX3_01	1.60
228	$NFKB_Q6_01	1.58
1957	Rectal Adenoma vs. Normal	1.54
165	NRSF	1.52
1317	$AP1_C	1.52

1855	NRSF	1.51

C. Developed Cells

# of promoters		Dev average
758	top758	5.83
190	UCSC_CpG-3+	2.38
64	znf354chighcon	1.92
1325	UCSC_CpG-2+	1.82
1135	UCSC_CpG-2	1.73
858	Prostate Carcinoma vs. Normal	1.68
1317	$AP1_C	1.64
165	NRSF	1.63
1957	Lymphoma - Top 10% Under-expressed (Wooster CellLine)	1.62
7476	cpg-1	1.62
860	Lung Adenocarcinoma vs. Normal	1.61
1407	$LHX3_01	1.58
1855	NRSF	1.57
2583	$TATA_01	1.53
2012	NFKB1-highcon	1.51

D. All Cell Average

# of promoters		tot average
758	top758	6.95
190	UCSC_CpG-3+	2.81
1325	UCSC_CpG-2+	1.97
64	znf354chighcon	1.85
1135	UCSC_CpG-2	1.83
1407	$LHX3_01	1.76
858	Prostate Carcinoma vs. Normal	1.75
1855	NRSF	1.73
1317	$AP1_C	1.68
1936	7gmd-highcon	1.63

7476	cpg-1	1.63
2012	NFKB1-highcon	1.63
1957	Lymphoma - Top 10% Under-expressed (Wooster CellLine)	1.62
2091	17md-con	1.62
1795	17md-highcon	1.61
487	SMTTTTGT	1.60
1826	motif1-highcon	1.60
860	Lung Adenocarcinoma vs. Normal	1.58
228	$NFKB_Q6_01	1.58
2727	7gmd-con	1.57
1957	Rectal Adenoma vs. Normal	1.56
2583	$TATA_01	1.56
165	NRSF	1.55
4194	NFKB1-con	1.54
5636	758m0highcon	1.54
1094	Small Cell Lung Carcinoma vs. Normal	1.52
232	znf354chigh	1.52
6378	TFAP2A-highcon	1.51
1572	$NFAT_Q4_01	1.50
1466	motif6-highcon	1.50

E. 758 genes that were present in at least half of the experiments (* represents p-value smaller than 10^{-15})

	# of promoters	**Top 758**	p-val				
758	top758	28.09		6065	nrf1-con	2.02	*
190	UCSC_CpG-3+	6.06	*	6199	10gm3-con	2.01	*
1325	UCSC_CpG-2+	3.60	*	3073	mat4-1cons	2.00	*
1135	UCSC_CpG-2	3.19	*	6217	10gm2-con	2.00	*
1407	$LHX3_01	2.70	*	2083	motif4-highcon	1.97	*
231	17gmeme	2.68	1.96E-05	7611	10md-highcon	1.97	*
487	SMTTTTGT	2.65	1.31E-09	1957	Colorectal Cancer - Top 10% Under-expressed (Wooster	1.95	1.76E-14

					CellLine)		
1936	7gmd-highcon	2.63	*	2389	tp53	1.95	*
2091	17md-con	2.57	*	3660	motif6-con	1.94	*
1317	$AP1_C	2.56	*	4514	nrf1-highcon	1.94	*
2012	NFKB1-highcon	2.55	*	8209	motif5-highcon	1.93	*
1826	motif1-highcon	2.54	*	1572	$NFAT_Q4_01	1.91	2.81E-11
2727	7gmd-con	2.52	*	858	Prostate Carcinoma vs. Normal	1.90	1.22E-06
1795	17md-highcon	2.52	*	1809	sef-1(cisgenome)	1.89	1.78E-12
1855	NRSF	2.48	*	5802	17ggm4-highcon	1.89	*
4194	NFKB1-con	2.36	*	3484	CREB1-highcon	1.89	*
451	$IRF_Q6	2.24	3.67E-06	3961	cpg_edge-1-1	1.89	*
138	yy1-highcon	2.24	6.59E-03	3918	p300highcon	1.87	*
5636	758m0highcon	2.24	*	5030	mp_3-1	1.85	*
228	$NFKB_Q6_01	2.22	8.58E-04	6370	motif3-con	1.84	*
64	znf354chighcon	2.19	5.13E-02	1297	LHX3-highcon	1.84	1.83E-08
6378	TFAP2A-highcon	2.18	*	8512	motif2-con	1.83	*
2875	motif3-highcon	2.17	*	2358	LHX3_cons-9	1.83	3.70E-14
5753	motif2-highcon	2.15	*	6519	NFKB1-high	1.83	*
3831	motif1-con	2.15	*	6519	nfkb1-high	1.83	*
1466	motif6-highcon	2.11	3.19E-14	4881	motif4-con	1.82	*
4325	10gm3-highcon	2.10	*	9095	pax4highcon	1.82	*
4326	10gm2-highcon	2.10	*	8711	sp1-highcon	1.82	*
3967	$SP1_Q6	2.10	*	8711	sp1-highcon	1.82	*
5600	10gmd3-highcon	2.10	*	1409	Egr-1	1.81	2.23E-08
2583	$TATA_01	2.10	*	9512	10md-con	1.81	*
5032	rest	2.07	*	1414	14md-con	1.81	1.32E-08
232	znf354chigh	2.06	5.09E-03	860	Lung Adenocarcinoma vs. Normal	1.80	1.03E-05
6330	7gmd2-highcon	2.04	*	9174	10gmd3-con	1.79	*
4285	MZF1-4-highcon	2.03	*	2632	deltaef1highcon	1.78	1.33E-14
5422	cpg_edge-3-2	1.78	*	10913	TFAP2A-high	1.57	*
1262	Breast Cancer - Top 10% Under-expressed (Shankavaram	1.78	3.48E-07	717	14md-highcon	1.57	2.44E-03

	CellLine)						
4167	lcg3+_gad_3	1.77	*	12262	sp1-con	1.57	*
3203	LHX3_cons-5	1.76	*	12262	sp1-con	1.57	*
7566	MZF1-4-con	1.76	*	1094	Small Cell Lung Carcinoma vs. Normal	1.57	1.50E-04
2700	LHX3-con	1.73	4.83E-13	12586	10md-high	1.56	*
1368	SOX9-highcon	1.72	3.26E-07	3413	ap1_c-highcon	1.56	2.30E-11
8530	nrf1-high	1.71	*	1725	osf2-highcon	1.56	3.14E-06
10514	HCG promoters	1.71	*	12123	mat4-2cons	1.55	*
8695	17ggm4-con	1.70	*	10963	motif2-high	1.55	*
1957	Rectal Adenoma vs. Normal	1.69	3.54E-09	671	lcg3+-5	1.55	2.46E-03
6059	motif1-high	1.68	*	8626	7gmd-high	1.54	*
4200	mp_7-0	1.67	*	7707	motif4-high	1.54	*
1513	Pax-9	1.67	4.70E-07	11790	nfkb1	1.53	*
1060	HEN1 (A)	1.67	3.91E-05	10116	ELK1-con	1.53	*
2230	exsr1-fli1	1.65	2.19E-09	165	NRSF	1.53	6.43E-02
1957	Lymphoma - Top 10% Under-expressed (Wooster CellLine)	1.64	4.06E-08	6503	znf354ccon	1.53	*
6058	ELK1-highcon	1.63	*	1328	$YY1_Q6	1.52	9.76E-05
3188	LHX3-3	1.62	2.49E-12	978	Invasive Ductal Breast Carcinoma vs. Normal	1.52	6.47E-04
260	yy1-high	1.62	2.18E-02	203	mat_0cons	1.52	5.06E-02
1530	Prostate Carcinoma vs. Normal	1.62	2.32E-06	925	Colon Adenocarcinoma vs. Normal	1.52	9.11E-04
7780	HMX3	1.61	*	1148	AR (A)	1.52	2.89E-04
11373	motif5-con	1.61	*	1777	Prostate Adenocarcinoma vs. Normal	1.50	1.59E-05
11029	Lcg_8	1.61	*	7445	ap1_c-con	1.50	*
5213	motif6-high	1.61	*	8870	motif3-high	1.50	*
8461	p300con	1.60	*				
6575	tlx1-nfic	1.60	*				

246	GCGNNANTTCC	1.60	2.71E-02
7449	CREB1-con	1.60	*
2384	lcg3+_gad_8	1.59	1.12E-08
4123	osf2-con	1.58	1.55E-14
11175	nrf1	1.58	*
11062	7gmd2-con	1.57	*
6055	rxra-vdr	1.57	*
12095	TFAP2A-con	1.57	*

F. Low CpG content CpG Islands (* represents p-value smaller than 10^{-15})

	# of promoters	**LCG**	p-val
3891	LCG promoters	5.77	
136	SEF-1	2.12	1.62E-07
1404	Pit-1	2.12	*
1462	PEBP	2.08	*
1420	Osf2	2.01	*
2583	$TATA_01	1.96	*
68	Brachyury	1.95	9.47E-04
1071	Lhx3(transf)	1.93	*
1446	AML	1.92	*
1139	core-binding factor	1.92	*
1404	IRF (A)	1.92	*
542	$ETS2_B	1.82	1.66E-14
1957	Rectal Adenoma vs. Normal	1.82	*
1407	$LHX3_01	1.81	*
1270	CHX10	1.78	*
1478	LEF1TCF1	1.78	*
925	Colon Adenocarcinoma vs. Normal	1.76	*
1481	Lmo2 complex	1.71	*
1317	$AP1_C	1.66	*
228	$NFKB_Q6_01	1.62	7.98E-05
1449	IRF-7	1.58	*
451	$IRF_Q6	1.57	9.44E-07
860	Lung Adenocarcinoma vs. Normal	1.56	6.21E-11
1094	Small Cell Lung	1.54	2.68E-13

1262	Carcinoma vs. Normal Squamous Cell Lung Carcinoma vs. Normal	1.51	1.41E-13

G. High CpG content CpG Islands (* represents p-value smaller than 10^{-15})

10514	HCG promoters	2.03	
758	top758	1.71	*
1250	Nrf-1	1.59	*
1409	Egr-1	1.56	*
886	CREB (a)	1.52	*
1414	DEAF1 (A)	1.52	*
1288	ATF (A)	1.51	*
3967	$SP1_Q6	1.50	*

Appendix Table 4.7. All motifs used in analysis and their origin

# of promoters	Motifs	Type	Program used
3073	mat4-1cons	Computational	CisGenome
3961	cpg_edge-1-1	Computational	CisGenome
203	mat_0cons	Computational	CisGenome
671	lcg3+-5	Computational	CisGenome
5422	cpg_edge-3-2	Computational	CisGenome
11029	Lcg_8	Computational	CisGenome
13791	lcg3+-4	Computational	CisGenome
14758	Lcg_5	Computational	CisGenome
12305	lcg3+-2	Computational	CisGenome
8071	cpg_edge-2-1	Computational	CisGenome
15199	Lcg_2	Computational	CisGenome
9424	Lcg_3	Computational	CisGenome
9531	lcg3+-9	Computational	CisGenome
10124	lcg3+-1	Computational	CisGenome
12903	Lcg_7	Computational	CisGenome
12144	758m3	Computational	CisGenome
11702	lcg3+-6	Computational	CisGenome
9405	lcg3+-0	Computational	CisGenome
1011	cpg_edge_3-1	Computational	CisGenome
12728	lcg3+-8-2	Computational	CisGenome
7847	lcg3+-8	Computational	CisGenome
7419	Lcg_1	Computational	CisGenome
5030	mp_3-1	Derived Motif	CisGenome
4200	mp_7-0	Derived Motif	CisGenome
11510	MP_1-1	Derived Motif	CisGenome
10355	mp_6-0	Derived Motif	CisGenome
8367	mp_4-0	Derived Motif	CisGenome
14463	mp_2-1	Derived Motif	CisGenome
14532	mp_5-0	Derived Motif	CisGenome
5032	rest	JASPAR	CisGenome
2230	exsr1-fli1	JASPAR	CisGenome
6575	tlx1-nfic	JASPAR	CisGenome
11175	nrf1	JASPAR	CisGenome
11790	nfkb1	JASPAR	CisGenome
6055	rxra-vdr	JASPAR	CisGenome
11662	e2f1	JASPAR	CisGenome
13099	nfkb	JASPAR	CisGenome
14916	nhlh1	JASPAR	CisGenome
12680	rela	JASPAR	CisGenome

13464	mizf	JASPAR	CisGenome
12530	myc-max	JASPAR	CisGenome
15113	myf	JASPAR	CisGenome
8517	CREB	JASPAR	CisGenome
4277	srf	JASPAR	CisGenome
12314	myb	JASPAR	CisGenome
13388	brca1	JASPAR	CisGenome
12307	nr3c1	JASPAR	CisGenome
13259	nfe2l2	JASPAR	CisGenome
13158	nr2f1	JASPAR	CisGenome
15990	rel	JASPAR	CisGenome
14537	nfya	JASPAR	CisGenome
10693	gata2	JASPAR	CisGenome
8152	rora2	JASPAR	CisGenome
16367	creb1	JASPAR	CisGenome
15854	foxc1	JASPAR	CisGenome
12898	irf1	JASPAR	CisGenome
18300	tfap2a	JASPAR	CisGenome
14721	nfatc2	JASPAR	CisGenome
17996	zeste	JASPAR	CisGenome
15011	yy1	JASPAR	CisGenome
19191	10gmd3	JASPAR	CisGenome
14700	rora1	JASPAR	CisGenome
11499	pbx1	JASPAR	CisGenome
17925	elk1	JASPAR	CisGenome
13381	hlf	JASPAR	CisGenome
15307	gata3	JASPAR	CisGenome
19332	sp1	JASPAR	CisGenome
16539	foxo3	JASPAR	CisGenome
17747	ap1	JASPAR	CisGenome
15790	sox9	JASPAR	CisGenome
15235	foxa1	JASPAR	CisGenome
13327	foxf2	JASPAR	CisGenome
16096	sry	JASPAR	CisGenome
14636	nkx3-1	JASPAR	CisGenome
10370	nfil3	JASPAR	CisGenome
8354	foxl1	JASPAR	CisGenome
1754	nr1h2-rxra	JASPAR	CisGenome
2358	LHX3_cons-9	non-PWM	CisGenome
3203	LHX3_cons-5	non-PWM	CisGenome
3188	LHX3-3	non-PWM	CisGenome
13379	mot_0-2_cons	non-PWM	CisGenome
9014	zeste_cons	non-PWM	CisGenome

351	myb_cons	non-PWM	CisGenome
10921	LHX3_cons-8	non-PWM	CisGenome
7862	LHX3_cons-2	non-PWM	CisGenome
5667	LHX3-2	non-PWM	CisGenome
10390	LHX3-1	non-PWM	CisGenome
1855	NRSF	Transfac	CisGenome
1809	sef-1(cisgenome)	Transfac	CisGenome
2389	tp53	Transfac	CisGenome
7780	HMX3	Transfac	CisGenome
8461	p300	Transfac	CisGenome
6503	znf354c	Transfac	CisGenome
8757	deltaef1	Transfac	CisGenome
10901	pax6	Transfac	CisGenome
14519	osf2	Transfac	CisGenome
12734	LMO2	Transfac	CisGenome
5893	CTCF	Transfac	CisGenome
13177	elk4	Transfac	CisGenome
7028	oct1	Transfac	CisGenome
3994	gklf	Transfac	CisGenome
16363	CREB1	Transfac	CisGenome
17337	ap1_c	Transfac	CisGenome
4167	lcg3+_gad_3	Computational	GADEM
2384	lcg3+_gad_8	Computational	GADEM
12521	lcg3+_gad_6	Computational	GADEM
12191	lcg3+_gad_1	Computational	GADEM
11299	lcg3+_gad_7	computational	GADEM
1896	lcg3+_gad_5	Computational	GADEM
11343	motif1	Computational	GADEM
14911	motif2	Computational	GADEM
12826	ctrl2	Computational	GADEM
11908	ctrl1	Computational	GADEM
14927	ctrl3	Computational	GADEM
11127	motif6	Computational	GADEM
14585	motif4	Computational	GADEM
16645	10gm3	Computational	GADEM
16659	10gm2	Computational	GADEM
16712	lcg_withcpg_gad_3	Computational	GADEM
15547	motif3	Computational	GADEM
17678	17ggm4	Computational	GADEM
18334	motif5	computational	GADEM
19535	lcg3+_gad_4	Computational	GADEM
9380	17md	Computational	Mdscan
10715	7gmd	Computational	Mdscan

14775	10md	Computational	Mdscan
6557	14md	Computational	Mdscan
19793	7gmd2	Computational	Mdscan
18595	1gmd2	Computational	Mdscan
18894	1gmd	Computational	Mdscan
1086	17gmeme	Computational	MEME
487	SMTTTTGT	motif derived	Oncomine
246	GCGNNANTTCC	motif derived	Oncomine
1407	$LHX3_01	Transfac	Oncomine
1317	$AP1_C	Transfac	Oncomine
228	$NFKB_Q6_01	Transfac	Oncomine
2583	$TATA_01	Transfac	Oncomine
165	NRSF	Transfac	Oncomine
1572	$NFAT_Q4_01	Transfac	Oncomine
3967	$SP1_Q6	Transfac	Oncomine
451	$IRF_Q6	Transfac	Oncomine
136	SEF-1	Transfac	Oncomine
1409	Egr-1	Transfac	Oncomine
1513	Pax-9	Transfac	Oncomine
1060	HEN1 (A)	Transfac	Oncomine
1530	MyoD	Transfac	Oncomine
1434	Hmx3	Transfac	Oncomine
542	$ETS2_B	Transfac	Oncomine
1148	AR (A)	Transfac	Oncomine
1478	LEF1TCF1	Transfac	Oncomine
1193	NF-kappaB (A)	Transfac	Oncomine
1292	NF-kappaB (C)	Transfac	Oncomine
1452	c-Rel	Transfac	Oncomine
1492	NF-kappaB (p65)	Transfac	Oncomine
1328	$YY1_Q6	Transfac	Oncomine
1381	NF-kappaB (D)	Transfac	Oncomine
1462	PEBP	Transfac	Oncomine
1446	AML	Transfac	Oncomine
1449	IRF-7	Transfac	Oncomine
1139	core-binding factor	Transfac	Oncomine
1420	Osf2	Transfac	Oncomine
68	Brachyury	Transfac	Oncomine
1481	Lmo2 complex	Transfac	Oncomine
1270	CHX10	Transfac	Oncomine
1294	N-Myc	Transfac	Oncomine
1404	IRF (A)	Transfac	Oncomine
1071	Lhx3(transf)	Transfac	Oncomine
1414	DEAF1 (A)	Transfac	Oncomine

1404	Pit-1	Transfac	Oncomine
1288	ATF (A)	Transfac	Oncomine
886	CREB (a)	Transfac	Oncomine
1250	Nrf-1	Transfac	Oncomine
1364	Genes with at least one ERa binding site within 20kb of transcriptional start site	Transfac	Oncomine

Chapter 5. References

1. *Statistics for NCBI Resources.* 2012. http://www.ncbi.nlm.nih.gov/About/tools/restable_stat_pubmed.html
2. *Genome News Network.* 2012. http://www.genomenewsnetwork.org/resources/sequenced_genomes/genome_guide_index.shtml
3. Vickaryous, M.K. and B.K. Hall, *Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest.* Biol Rev Camb Philos Soc, 2006. **81**(3): p. 425-55.
4. Larranaga, P., et al., *Machine learning in bioinformatics.* Brief Bioinform, 2006. **7**(1): p. 86-112.
5. Sousa, F.L., et al., *A bioinformatics classifier and database for heme-copper oxygen reductases.* PLoS One, 2011. **6**(4): p. e19117.
6. Fu, J., et al., *Partial least squares regression, support vector machine regression, and transcriptome-based distances for prediction of maize hybrid performance with gene expression data.* Theor Appl Genet, 2012. **124**(5): p. 825-33.
7. Yang, Z., *PAML: a program package for phylogenetic analysis by maximum likelihood.* Comput Appl Biosci, 1997. **13**(5): p. 555-6.
8. Li, L., *GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery.* J Comput Biol, 2009. **16**(2): p. 317-29.
9. Vapnik, C.C.a.V., *Support-Vector Networks.* MACHINE LEARNING, 1995. **20**(3): p. 273-297.
10. Krystal, M., et al., *Evolution of influenza A and B viruses: conservation of structural features in the hemagglutinin genes.* Proc Natl Acad Sci U S A, 1982. **79**(15): p. 4800-4.
11. Rota, P.A., et al., *Cocirculation of two distinct evolutionary lineages of influenza type B virus since 1983.* Virology, 1990. **175**(1): p. 59-68.
12. Shaw, M.W., et al., *Reappearance and global spread of variants of influenza B/Victoria/2/87 lineage viruses in the 2000-2001 and 2001-2002 seasons.* Virology, 2002. **303**(1): p. 1-8.
13. Kanegae, Y., et al., *Evolutionary pattern of the hemagglutinin gene of influenza B viruses isolated in Japan: cocirculating lineages in the same epidemic season.* J Virol, 1990. **64**(6): p. 2860-5.
14. Verhoeyen, M., et al., *Complete nucleotide sequence of the influenza B/Singapore/222/79 virus hemagglutinin gene and comparison with the B/Lee/40 hemagglutinin.* Nucleic Acids Res, 1983. **11**(14): p. 4703-12.
15. Berton, M.T., C.W. Naeve, and R.G. Webster, *Antigenic structure of the influenza B virus hemagglutinin: nucleotide sequence analysis of antigenic variants selected with monoclonal antibodies.* J Virol, 1984. **52**(3): p. 919-27.
16. Berton, M.T. and R.G. Webster, *The antigenic structure of the influenza B virus hemagglutinin: operational and topological mapping with monoclonal antibodies.* Virology, 1985. **143**(2): p. 583-94.

17. Hovanec, D.L. and G.M. Air, *Antigenic structure of the hemagglutinin of influenza virus B/Hong Kong/8/73 as determined from gene sequence analysis of variants selected with monoclonal antibodies.* Virology, 1984. **139**(2): p. 384-92.
18. Webster, R.G. and M.T. Berton, *Analysis of antigenic drift in the haemagglutinin molecule of influenza B virus with monoclonal antibodies.* J Gen Virol, 1981. **54**(Pt 2): p. 243-51.
19. Krystal, M., et al., *Sequential mutations in hemagglutinins of influenza B virus isolates: definition of antigenic domains.* Proc Natl Acad Sci U S A, 1983. **80**(14): p. 4527-31.
20. Bootman, J.S. and J.S. Robertson, *Sequence analysis of the hemagglutinin of B/Ann Arbor/1/86, an epidemiologically significant variant of influenza B virus.* Virology, 1988. **166**(1): p. 271-4.
21. Rota, P.A., et al., *Antigenic and genetic characterization of the haemagglutinins of recent cocirculating strains of influenza B virus.* J Gen Virol, 1992. **73 (Pt 10)**: p. 2737-42.
22. Lubeck, M.D., J.L. Schulman, and P. Palese, *Antigenic variants of influenza viruses: marked differences in the frequencies of variants selected with different monoclonal antibodies.* Virology, 1980. **102**(2): p. 458-62.
23. Air, G.M., et al., *Evolutionary changes in influenza B are not primarily governed by antibody selection.* Proc Natl Acad Sci U S A, 1990. **87**(10): p. 3884-8.
24. Pechirra, P., et al., *Molecular characterization of the HA gene of influenza type B viruses.* J Med Virol, 2005. **77**(4): p. 541-9.
25. Yang, Z., et al., *Codon-substitution models for heterogeneous selection pressure at amino acid sites.* Genetics, 2000. **155**(1): p. 431-49.
26. Nakagawa, N., et al., *Neutralizing epitopes specific for influenza B virus Yamagata group strains are in the 'loop'.* J Gen Virol, 2003. **84**(Pt 4): p. 769-73.
27. Nakagawa, N., et al., *Antigenic variants with amino acid deletions clarify a neutralizing epitope specific for influenza B virus Victoria group strains.* J Gen Virol, 2001. **82**(Pt 9): p. 2169-72.
28. Nakagawa, N., R. Kubota, and Y. Okuno, *Variation of the conserved neutralizing epitope in influenza B virus victoria group isolates in Japan.* J Clin Microbiol, 2005. **43**(8): p. 4212-4.
29. Nakagawa, N., et al., *Characterization of new epidemic strains of influenza B virus by using neutralizing monoclonal antibodies.* J Med Virol, 2001. **65**(4): p. 745-50.
30. Wang, Q., et al., *Crystal Structure of Unliganded Influenza B Virus Hemagglutinin.* J Virol, 2008. **82**: p. 3011-20.
31. Chen, R. and E.C. Holmes, *The Evolutionary Dynamics of Human Influenza B Virus.* J Mol Evol, 2008.
32. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.
33. Macken, C., et al., *The value of a database in surveillance and Vaccine selection*, in *Options for the control of the influenza IV*, A. Osterhaus, N. Cox, and A.W. Hampson, Editors. 2001, Elsevier Sciences: Amsterdam. p. 103-106.
34. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence*

weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 1994. **22**(22): p. 4673-80.

35. Yang, Z., W.S. Wong, and R. Nielsen, *Bayes empirical bayes inference of amino acid sites under positive selection.* Mol Biol Evol, 2005. **22**(4): p. 1107-18.
36. Deely, J.J. and D.V. Lindley, *Bayes empirical bayes.* J. Am. Stat. Assoc., 1981. **76**: p. 833-841.
37. Nerome, R., et al., *Evolutionary characteristics of influenza B virus since its first isolation in 1940: dynamic circulation of deletion and insertion mechanism.* Arch Virol, 1998. **143**(8): p. 1569-83.
38. Matsuzaki, Y., et al., *Genetic diversity of influenza B virus: the frequent reassortment and cocirculation of the genetically distinct reassortant viruses in a community.* J Med Virol, 2004. **74**(1): p. 132-40.
39. McCullers, J.A., T. Saito, and A.R. Iverson, *Multiple genotypes of influenza B virus circulated between 1979 and 2003.* J Virol, 2004. **78**(23): p. 12817-28.
40. Chen, J.M., et al., *Exploration of the emergence of the Victoria lineage of influenza B virus.* Arch Virol, 2007. **152**(2): p. 415-22.
41. Bush, R.M., et al., *Positive selection on the H3 hemagglutinin gene of human influenza virus A.* Mol Biol Evol, 1999. **16**(11): p. 1457-65.
42. Anisimova, M., J.P. Bielawski, and Z. Yang, *Accuracy and power of bayes prediction of amino acid sites under positive selection.* Mol Biol Evol, 2002. **19**(6): p. 950-8.
43. Abed, Y., et al., *Evolution of surface and nonstructural-1 genes of influenza B viruses isolated in the Province of Quebec, Canada, during the 1998-2001 period.* Virus Genes, 2003. **27**(2): p. 125-35.
44. Lugovtsev, V.Y., G.M. Vodeiko, and R.A. Levandowski, *Mutational pattern of influenza B viruses adapted to high growth replication in embryonated eggs.* Virus Res, 2005. **109**(2): p. 149-57.
45. Oxford, J.S., et al., *A host-cell-selected variant of influenza B virus with a single nucleotide substitution in HA affecting a potential glycosylation site was attenuated in virulence for volunteers.* Arch Virol, 1990. **110**(1-2): p. 37-46.
46. Oxford, J.S., et al., *Direct isolation in eggs of influenza A (H1N1) and B viruses with haemagglutinins of different antigenic and amino acid composition.* J Gen Virol, 1991. **72 (Pt 1)**: p. 185-9.
47. Lugovtsev, V.Y., et al., *Generation of the influenza B viruses with improved growth phenotype by substitution of specific amino acids of hemagglutinin.* Virology, 2007. **365**(2): p. 315-23.
48. McCullers, J.A., et al., *Reassortment and insertion-deletion are strategies for the evolution of influenza B viruses in nature.* J Virol, 1999. **73**(9): p. 7343-8.
49. Schulze, I.T., *Effects of glycosylation on the properties and functions of influenza virus hemagglutinin.* J Infect Dis, 1997. **176 Suppl 1**: p. S24-8.
50. Skehel, J.J., et al., *A carbohydrate side chain on hemagglutinins of Hong Kong influenza viruses inhibits recognition by a monoclonal antibody.* Proc Natl Acad Sci U S A, 1984. **81**(6): p. 1779-83.
51. Skehel, J.J. and D.C. Wiley, *Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin.* Annu Rev Biochem, 2000. **69**: p. 531-69.

52. Caton, A.J., et al., *The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype).* Cell, 1982. **31**(2 Pt 1): p. 417-27.
53. Daniels, R.S., et al., *Analyses of the antigenicity of influenza haemagglutinin at the pH optimum for virus-mediated membrane fusion.* J Gen Virol, 1983. **64 (Pt 8)**: p. 1657-62.
54. Nakagawa, N., et al., *Influenza B virus victoria group with a new glycosylation site was epidemic in Japan in the 2002-2003 season.* J Clin Microbiol, 2004. **42**(7): p. 3295-7.
55. Schild, G.C., et al., *Evidence for host-cell selection of influenza virus antigenic variants.* Nature, 1983. **303**(5919): p. 706-709.
56. Robertson, J.S., et al., *Alterations in the hemagglutinin associated with adaptation of influenza B virus to growth in eggs.* Virology, 1985. **143**(1): p. 166-74.
57. Gambaryan, A.S., J.S. Robertson, and M.N. Matrosovich, *Effects of egg-adaptation on the receptor-binding properties of human influenza A and B viruses.* Virology, 1999. **258**(2): p. 232-9.
58. Saito, T., et al., *Antigenic alteration of influenza B virus associated with loss of a glycosylation site due to host-cell adaptation.* J Med Virol, 2004. **74**(2): p. 336-43.
59. Robertson, J.S., et al., *The hemagglutinin of influenza B virus present in clinical material is a single species identical to that of mammalian cell-grown virus.* Virology, 1990. **179**(1): p. 35-40.
60. Ikonen, N., et al., *Reappearance of influenza B/Victoria/2/87-lineage viruses: epidemic activity, genetic diversity and vaccination efficacy in the Finnish Defence Forces.* Epidemiol Infect, 2005. **133**(2): p. 263-71.
61. Nakagawa, N., et al., *Heterogeneity of influenza B virus strains in one epidemic season differentiated by monoclonal antibodies and nucleotide sequences.* J Clin Microbiol, 2000. **38**(9): p. 3467-9.
62. Muyanga, J., et al., *Antigenic and genetic analyses of influenza B viruses isolated in Lusaka, Zambia in 1999.* Arch Virol, 2001. **146**(9): p. 1667-79.
63. Nakagawa, N., et al., *Discovery of the neutralizing epitope common to influenza B virus victoria group isolates in Japan.* J Clin Microbiol, 2006. **44**(4): p. 1564-6.
64. Knossow, M., et al., *Mechanism of neutralization of influenza virus infectivity by antibodies.* Virology, 2002. **302**(2): p. 294-8.
65. Barbey-Martin, C., et al., *An antibody that prevents the hemagglutinin low pH fusogenic transition.* Virology, 2002. **294**(1): p. 70-4.
66. Bizebard, T., et al., *Structural studies on viral escape from antibody neutralization.* Curr Top Microbiol Immunol, 2001. **260**: p. 55-64.
67. Gigant, B., et al., *A neutralizing antibody Fab-influenza haemagglutinin complex with an unprecedented 2:1 stoichiometry: characterization and crystallization.* Acta Crystallogr D Biol Crystallogr, 2000. **56 (Pt 8)**: p. 1067-9.
68. Bizebard, T., et al., *Structure of influenza virus haemagglutinin complexed with a neutralizing antibody.* Nature, 1995. **376**(6535): p. 92-94.
69. Fleury, D., et al., *A complex of influenza hemagglutinin with a neutralizing antibody that binds outside the virus receptor binding site.* Nat Struct Biol, 1999. **6**(6): p. 530-4.

70. Shitani, M., et al., *Genome-wide analysis of DNA methylation identifies novel cancer-related genes in hepatocellular carcinoma.* Tumour Biol, 2012.
71. Nayak, V., C. Xu, and J. Min, *Composition, recruitment and regulation of the PRC2 complex.* Nucleus, 2011. **2**(4).
72. Squazzo, S.L., et al., *Suz12 binds to silenced regions of the genome in a cell-type-specific manner.* Genome Res, 2006. **16**(7): p. 890-900.
73. Lanzuolo, C., et al., *Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex.* Nat Cell Biol, 2007. **9**(10): p. 1167-74.
74. Ringrose, L. and R. Paro, *Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins.* Annu Rev Genet, 2004. **38**: p. 413-43.
75. Schwartz, Y.B., et al., *Genome-wide analysis of Polycomb targets in Drosophila melanogaster.* Nat Genet, 2006. **38**(6): p. 700-5.
76. Ku, M., et al., *Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains.* PLoS Genet, 2008. **4**(10): p. e1000242.
77. Bracken, A.P., et al., *Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions.* Genes Dev, 2006. **20**(9): p. 1123-36.
78. Lee, T.I., et al., *Control of developmental regulators by Polycomb in human embryonic stem cells.* Cell, 2006. **125**(2): p. 301-13.
79. Boyer, L.A., et al., *Polycomb complexes repress developmental regulators in murine embryonic stem cells.* Nature, 2006. **441**(7091): p. 349-53.
80. Sing, A., et al., *A vertebrate Polycomb response element governs segmentation of the posterior hindbrain.* Cell, 2009. **138**(5): p. 885-97.
81. Pan, G., et al., *Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells.* Cell Stem Cell, 2007. **1**(3): p. 299-312.
82. Zhao, X.D., et al., *Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells.* Cell Stem Cell, 2007. **1**(3): p. 286-98.
83. Choi, J.H., et al., *Genome-wide DNA methylation maps in follicular lymphoma cells determined by methylation-enriched bisulfite sequencing.* PLoS One, 2010. **5**(9).
84. Kwong, C., et al., *Stability and dynamics of polycomb target sites in Drosophila development.* PLoS Genet, 2008. **4**(9): p. e1000178.
85. Schuettengruber, B., et al., *Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos.* PLoS Biol, 2009. **7**(1): p. e13.
86. Tolhuis, B., et al., *Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster.* Nat Genet, 2006. **38**(6): p. 694-9.
87. Ke, X.S., et al., *Genome-wide profiling of histone h3 lysine 4 and lysine 27 trimethylation reveals an epigenetic signature in prostate carcinogenesis.* PLoS One, 2009. **4**(3): p. e4687.
88. Chan, C.S., L. Rastelli, and V. Pirrotta, *A Polycomb response element in the Ubx gene that determines an epigenetically inherited state of repression.* EMBO J, 1994. **13**(11): p. 2553-64.
89. Mihaly, J., R.K. Mishra, and F. Karch, *A Conserved Sequence Motif in Polycomb-Response Elements.* Molecular Cell, 1998. **1**: p. 1065-1066.

90. Schuettengruber, B., et al., *Genome Regulation by Polycomb and Trithorax Proteins.* Cell, 2007. **128**: p. 735-745.
91. Nguyen, N., et al., *Molecular cloning and functional characterization of the transcription factor YY2.* J Biol Chem, 2004. **279**(24): p. 25927-34.
92. Brown, J.L., et al., *The Drosophila Polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1.* Mol Cell, 1998. **1**(7): p. 1057-64.
93. Wang, L., et al., *Hierarchical recruitment of polycomb group silencing complexes.* Mol Cell, 2004. **14**(5): p. 637-46.
94. Strutt, H., G. Cavalli, and R. Paro, *Co-localization of Polycomb protein and GAGA factor on regulatory elements responsible for the maintenance of homeotic gene expression.* EMBO J, 1997. **16**(12): p. 3621-32.
95. Decoville, M., et al., *DSP1, an HMG-like Protein, Is Involved in the Regulation of Homeotic Genes.* Genetics, 2001. **157**(1): p. 237-244.
96. Hodgson, J.W., B. Argiropoulos, and H.W. Brock, *Site-specific recognition of a 70-base-pair element containing d(GA)(n) repeats mediates bithoraxoid polycomb group response element-dependent silencing.* Mol Cell Biol, 2001. **21**(14): p. 4528-43.
97. Dejardin, J. and G. Cavalli, *Chromatin inheritance upon Zeste-mediated Brahma recruitment at a minimal cellular memory module.* EMBO J, 2004. **23**(4): p. 857-868.
98. Hagstrom, K., M. Muller, and P. Schedl, *A Polycomb and GAGA dependent silencer adjoins the Fab-7 boundary in the Drosophila bithorax complex.* Genetics, 1997. **146**(4): p. 1365-80.
99. Brickman, J.M., M. Adam, and M. Ptashne, *Interactions between an HMG-1 protein and members of the Rel family.* Proc Natl Acad Sci U S A, 1999. **96**(19): p. 10679-83.
100. Dejardin, J., et al., *Recruitment of Drosophila Polycomb group proteins to chromatin by DSP1.* Nature, 2005. **434**(7032): p. 533-8.
101. Blastyak, A., et al., *Efficient and specific targeting of Polycomb group proteins requires cooperative interaction between Grainyhead and Pleiohomeotic.* Mol Cell Biol, 2006. **26**(4): p. 1434-44.
102. Brown, J.L., et al., *An Sp1/KLF binding site is important for the activity of a Polycomb group response element from the Drosophila engrailed gene.* Nucleic Acids Research, 2005. **33**(16): p. 5181–5189.
103. Ringrose, L. and R. Paro, *Polycomb/Trithorax response elements and epigenetic memory of cell identity.* Development, 2007. **134**(2): p. 223-32.
104. Kassis, J.A., *Unusual properties of regulatory DNA from the Drosophila engrailed gene: three "pairing-sensitive" sites within a 1.6-kb region.* Genetics, 1994. **136**(3): p. 1025-38.
105. Gindhart, J.G., Jr. and T.C. Kaufman, *Identification of Polycomb and trithorax group responsive elements in the regulatory region of the Drosophila homeotic gene Sex combs reduced.* Genetics, 1995. **139**(2): p. 797-814.
106. Americo, J., et al., *A complex array of DNA-binding proteins required for pairing-sensitive silencing by a polycomb group response element from the Drosophila engrailed gene.* Genetics, 2002. **160**(4): p. 1561-71.

107. Bloyer, S., et al., *Identification and characterization of polyhomeotic PREs and TREs.* Dev Biol, 2003. **261**(2): p. 426-42.
108. Gruzdeva, N., et al., *The Mcp element from the bithorax complex contains an insulator that is capable of pairwise interactions and can facilitate enhancer-promoter communication.* Mol Cell Biol, 2005. **25**(9): p. 3682-9.
109. DeVido, S.K., et al., *The role of Polycomb-group response elements in regulation of engrailed transcription in Drosophila.* Development, 2008. **135**(4): p. 669-76.
110. Kozma, G., W. Bender, and L. Sipos, *Replacement of a Drosophila Polycomb response element core, and in situ analysis of its DNA motifs.* Mol Genet Genomics, 2008. **279**(6): p. 595-603.
111. Cunningham, M.D., J.L. Brown, and J.A. Kassis, *Characterization of the polycomb group response elements of the Drosophila melanogaster invected Locus.* Mol Cell Biol, 2010. **30**(3): p. 820-8.
112. Horard, B., et al., *Structure of a polycomb response element and in vitro binding of polycomb group complexes containing GAGA factor.* Mol Cell Biol, 2000. **20**(9): p. 3187-97.
113. Ringrose, L., et al., *Genome-wide prediction of Polycomb/Trithorax response elements in Drosophila melanogaster.* Dev Cell, 2003. **5**(5): p. 759-71.
114. Woo, C.J., et al., *A region of the human HOXD cluster that confers polycomb-group responsiveness.* Cell, 2010. **140**(1): p. 99-110.
115. Mikkelsen, T.S., et al., *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**: p. 553-560.
116. Zhang, L., et al., *Genome-wide analysis of histone H3 lysine 27 trimethylation by ChIP-chip in gastric cancer patients.* J Gastroenterol, 2009. **44**(4): p. 305-12.
117. Araki, Y., et al., *Genome-wide analysis of histone methylation reveals chromatin state-based regulation of gene transcription and function of memory CD8+ T cells.* Immunity, 2009. **30**(6): p. 912-25.
118. Miao, F. and R. Natarajan, *Mapping Global Histone Methylation Patterns in the Coding Regions of Human Genes.* Molecular and Cellular Biology, 2005. **25**(11): p. 4650-4661.
119. Kondo, Y., et al., *Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation.* Nat Genet, 2008. **40**(6): p. 741-50.
120. Pasini, D., et al., *The polycomb group protein Suz12 is required for embryonic stem cell differentiation.* Mol Cell Biol, 2007. **27**(10): p. 3769-79.
121. Kirmizis, A., et al., *Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27.* Genes Dev, 2004. **18**(13): p. 1592-605.
122. Wei, G., et al., *Global mapping of H3K4me3 and H3K27me3 reveals specificity and plasticity in lineage fate determination of differentiating CD4+ T cells.* Immunity, 2009. **30**(1): p. 155-67.
123. Mikkelsen, T.S., et al., *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**(7153): p. 553-60.
124. Fiedler, T. and M. Rehmsmeier, *jPREdictor: a versatile tool for the prediction of cis-regulatory elements.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W546-50.
125. Hauenschild, A., et al., *Evolutionary plasticity of polycomb/trithorax response elements in Drosophila species.* PLoS Biol, 2008. **6**(10): p. e261.

126. Liu, Y., Z. Shao, and G.C. Yuan, *Prediction of Polycomb target genes in mouse embryonic stem cells.* Genomics, 2010.
127. Sandelin, A., et al., *JASPAR: an open-access database for eukaryotic transcription factor binding profiles.* Nucl. Acids Res., 2004. **32**(suppl_1): p. D91-94.
128. Matys, V., et al., *TRANSFAC(R): transcriptional regulation, from patterns to profiles.* Nucl. Acids Res., 2003. **31**(1): p. 374-378.
129. Rhodes, D.R., et al., *ONCOMINE: a cancer microarray database and integrated data-mining platform.* Neoplasia, 2004. **6**(1): p. 1-6.
130. van Steensel, B., J. Delrow, and H.J. Bussemaker, *Genomewide analysis of Drosophila GAGA factor target genes reveals context-dependent DNA binding.* Proc Natl Acad Sci U S A, 2003. **100**(5): p. 2580-5.
131. Guyon, I., et al., *Gene Selection for Cancer Classification using Support Vector Machines.* Mach. Learn., 2002. **46**(1-3): p. 389-422.
132. Tang, E.K., P.N. Suganthan, and X. Yao, *Gene selection algorithms for microarray data based on least squares support vector machine.* BMC Bioinformatics, 2006. **7**: p. 95.
133. Zhang, S.W., et al., *Classification of protein quaternary structure with support vector machine.* Bioinformatics, 2003. **19**(18): p. 2390-6.
134. Shamim, M.T., M. Anwaruddin, and H.A. Nagarajaram, *Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs.* Bioinformatics, 2007. **23**(24): p. 3320-7.
135. Leslie, C., E. Eskin, and W.S. Noble, *The spectrum kernel: a string kernel for SVM protein classification.* Pac Symp Biocomput, 2002: p. 564-75.
136. Furey, T.S., et al., *Support vector machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 2000. **16**(10): p. 906-14.
137. Laufer, S. and B. Rubinsky, *Tissue characterization with an electrical spectroscopy SVM classifier.* IEEE Trans Biomed Eng, 2009. **56**(2): p. 525-8.
138. McQuisten, K.A. and A.S. Peek, *Comparing artificial neural networks, general linear models and support vector machines in building predictive models for small interfering RNAs.* PLoS One, 2009. **4**(10): p. e7522.
139. Shen, L. and E.C. Tan, *Reducing multiclass cancer classification to binary by output coding and SVM.* Comput Biol Chem, 2006. **30**(1): p. 63-71.
140. Das, R., et al., *Computational prediction of methylation status in human genomic sequences.* Proc Natl Acad Sci U S A, 2006. **103**(28): p. 10713-6.
141. Bock, C., et al., *CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure.* PLoS Genet, 2006. **2**(3): p. e26.
142. Tanay, A., et al., *Hyperconserved CpG domains underlie Polycomb-binding sites.* Proc Natl Acad Sci U S A, 2007. **104**(13): p. 5521-6.
143. Mendenhall, E.M., et al., *GC-rich sequence elements recruit PRC2 in mammalian ES cells.* PLoS Genet, 2010. **6**(12): p. e1001244.
144. Meissner, A., et al., *Genome-scale DNA methylation maps of pluripotent and differentiated cells.* Nature, 2008. **454**(7205): p. 766-70.

145. Flybase. *A database of Drosophila genes and genomes.* 2010; Available from: http://flybase.org.
146. Li, J.B., et al., *Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene.* Cell, 2004. **117**(4): p. 541-52.
147. Bernstein, B.E., et al., *Genomic maps and comparative analysis of histone modifications in human and mouse.* Cell, 2005. **120**(2): p. 169-81.
148. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.
149. Orlando, V., H. Strutt, and R. Paro, *Analysis of chromatin structure by in vivo formaldehyde cross-linking.* Methods, 1997. **11**(2): p. 205-14.
150. Chen, X., et al., *W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data.* Bioinformatics, 2008. **24**(9): p. 1121-8.
151. Machanick, P. and T.L. Bailey, *MEME-ChIP: motif analysis of large DNA datasets.* Bioinformatics, 2011. **27**(12): p. 1696-7.
152. Jiang, H., et al., *CisGenome Browser: a flexible tool for genomic data visualization.* Bioinformatics, 2010. **26**(14): p. 1781-2.
153. Schlesinger, Y., et al., *Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer.* Nat Genet, 2007. **39**(2): p. 232-6.
154. Gal-Yam, E.N., et al., *Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line.* Proceedings of the National Academy of Sciences, 2008. **105**(35): p. 12979-12984.
155. Choi, J.-H., et al., *Genome-Wide DNA Methylation Maps in Follicular Lymphoma Cells Determined by Methylation-Enriched Bisulfite Sequencing.* PLoS One, 2010. **5**(9): p. e13020.
156. Ke, X.S., et al., *Global profiling of histone and DNA methylation reveals epigenetic-based regulation of gene expression during epithelial to mesenchymal transition in prostate cells.* BMC Genomics, 2010. **11**: p. 669.
157. Tserel, L., et al., *Genome-wide promoter analysis of histone modifications in human monocyte-derived antigen presenting cells.* BMC Genomics, 2010. **11**: p. 642.
158. Chang, C. and C. Lin. *LIBSVM: a library for support vector machines.* 2001; Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm.
159. Chipman, H.A., E.I. George, and R.E. McCulloch, *BART: Bayesian Additive Rgression Trees.* The Annuals of Applied Statistics, 2010. **4**(1): p. 266-98.
160. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
161. Iwaki, D.D. and J.A. Lengyel, *A Delta-Notch signaling border regulated by Engrailed/Invected repression specifies boundary cells in the Drosophila hindgut.* Mech Dev, 2002. **114**(1-2): p. 71-84.
162. Pourreyron, C., et al., *Wnt5a Is Strongly Expressed at the Leading Edge in Non-Melanoma Skin Cancer, Forming Active Gradients, while Canonical Wnt Signalling Is Repressed.* PLoS One, 2012. **7**(2): p. e31827.

163. Jones, W.M. and A. Bejsovec, *RacGap50C negatively regulates wingless pathway activity during Drosophila embryonic development.* Genetics, 2005. **169**(4): p. 2075-86.
164. Billin, A.N., K.A. Cockerill, and S.J. Poole, *Isolation of a family of Drosophila POU domain genes expressed in early development.* Mech Dev, 1991. **34**(2-3): p. 75-84.
165. Lloyd, A. and S. Sakonju, *Characterization of two Drosophila POU domain genes, related to oct-1 and oct-2, and the regulation of their expression patterns.* Mech Dev, 1991. **36**(1-2): p. 87-102.
166. Dick, T., et al., *Two closely linked Drosophila POU domain genes are expressed in neuroblasts and sensory elements.* Proc Natl Acad Sci U S A, 1991. **88**(17): p. 7645-9.
167. Poole, S.J., *Conservation of complex expression domains of the pdm-2 POU domain gene between Drosophila virilis and Drosophila melanogaster.* Mech Dev, 1995. **49**(1-2): p. 107-16.
168. Chen, R., et al., *Dachshund and eyes absent proteins form a complex and function synergistically to induce ectopic eye development in Drosophila.* Cell, 1997. **91**(7): p. 893-903.
169. Jimenez, G., C.P. Verrijzer, and D. Ish-Horowicz, *A conserved motif in goosecoid mediates groucho-dependent repression in Drosophila embryos.* Mol Cell Biol, 1999. **19**(3): p. 2080-7.
170. de Navascues, J. and J. Modolell, *tailup, a LIM-HD gene, and Iro-C cooperate in Drosophila dorsal mesothorax specification.* Development, 2007. **134**(9): p. 1779-88.
171. Mann, T., R. Bodmer, and P. Pandur, *The Drosophila homolog of vertebrate Islet1 is a key component in early cardiogenesis.* Development, 2009. **136**(2): p. 317-26.
172. Lie, D.C., et al., *Wnt signalling regulates adult hippocampal neurogenesis.* Nature, 2005. **437**(7063): p. 1370-5.
173. Strutt, H. and R. Paro, *The polycomb group protein complex of Drosophila melanogaster has different compositions at different target genes.* Mol Cell Biol, 1997. **17**(12): p. 6773-83.
174. Lemons, D. and W. McGinnis, *Genomic evolution of Hox gene clusters.* Science, 2006. **313**(5795): p. 1918-22.
175. Petruk, S., et al., *Transcription of bxd noncoding RNAs promoted by trithorax represses Ubx in cis by transcriptional interference.* Cell, 2006. **127**(6): p. 1209-21.
176. Sanders, L.R., M. Patel, and J.W. Mahaffey, *The Drosophila gap gene giant has an anterior segment identity function mediated through disconnected and teashirt.* Genetics, 2008. **179**(1): p. 441-53.
177. Wagner-Bernholz, J.T., et al., *Identification of target genes of the homeotic gene Antennapedia by enhancer detection.* Genes Dev, 1991. **5**(12B): p. 2467-80.
178. Biggin, M.D. and R. Tjian, *A purified Drosophila homeodomain protein represses transcription in vitro.* Cell, 1989. **58**(3): p. 433-40.
179. Williams, T.M., et al., *The regulation and evolution of a genetic switch controlling sexually dimorphic traits in Drosophila.* Cell, 2008. **134**(4): p. 610-23.

180. Tabuchi, K., et al., *A novel Drosophila paired-like homeobox gene related to Caenorhabditis elegans unc-4 is expressed in subsets of postmitotic neurons and epidermal cells.* Neurosci Lett, 1998. **257**(1): p. 49-52.
181. Singh, A. and K.W. Choi, *Initial state of the Drosophila eye before dorsoventral specification is equivalent to ventral.* Development, 2003. **130**(25): p. 6351-60.
182. Luque, C.M. and M. Milan, *Growth control in the proliferative region of the Drosophila eye-head primordium: the elbow-noc gene complex.* Dev Biol, 2007. **301**(2): p. 327-39.
183. Lin, W.H., et al., *Expression of a Drosophila GATA transcription factor in multiple tissues in the developing embryos. Identification of homozygous lethal mutants with P-element insertion at the promoter region.* J Biol Chem, 1995. **270**(42): p. 25150-8.
184. Senger, K., K. Harris, and M. Levine, *GATA factors participate in tissue-specific immune responses in Drosophila larvae.* Proc Natl Acad Sci U S A, 2006. **103**(43): p. 15957-62.
185. Lai, Z.C., M.E. Fortini, and G.M. Rubin, *The embryonic expression patterns of zfh-1 and zfh-2, two Drosophila genes encoding novel zinc-finger homeodomain proteins.* Mech Dev, 1991. **34**(2-3): p. 123-34.
186. Saze, H., et al., *DNA Methylation in Plants: Relationship with Small RNAs and Histone Modifications, and Functions in Transposon Inactivation.* Plant Cell Physiol, 2012.
187. Simons, C., et al., *Transposon-free regions in mammalian genomes.* Genome Res, 2006. **16**(2): p. 164-72.
188. Bernstein, B.E., et al., *A bivalent chromatin structure marks key developmental genes in embryonic stem cells.* Cell, 2006. **125**(2): p. 315-26.
189. Aichinger, E., et al., *CHD3 proteins and polycomb group proteins antagonistically determine cell identity in Arabidopsis.* PLoS Genet, 2009. **5**(8): p. e1000605.
190. Schwartz, Y.B., et al., *Alternative epigenetic chromatin states of polycomb target genes.* PLoS Genet, 2010. **6**(1): p. e1000805.
191. Matharu, N.K., et al., *Vertebrate Homologue of Drosophila GAGA Factor.* J Mol Biol, 2010.
192. Kent WJ, S.C., Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D, *The human genome browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.
193. Bailey, T.L., et al., *MEME: discovering and analyzing DNA and protein sequence motifs.* Nucl. Acids Res., 2006. **34**(suppl_2): p. W369-373.
194. Liu, X., D. Brutlag, and J. Liu, *An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments.* Nat Biotechnol, 2002. **20**(8): p. 825-839.
195. Chen, X., et al., *W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data.* Bioinformatics, 2008. **24**(9): p. 1121-1128.
196. Li, L., *GADEM: A Genetic Algorithm Guided Formation of Spaced Dyads Coupled with an EM Algorithm for Motif Discovery.* Journal of Computational Biology, 2009. **16**(2): p. 317-329.

197. Pavesi, G., et al., *Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes.* Nucl. Acids Res., 2004. **32**(suppl_2): p. W199-203.
198. Sandelin, A., et al., *JASPAR: an open-access database for eukaryotic transcription factor binding profiles.* Nucleic Acids Res, 2004. **32**(Database issue): p. D91-4.
199. Wingender, E., et al., *TRANSFAC: a database on transcription factors and their DNA binding sites.* Nucleic Acids Res, 1996. **24**(1): p. 238-41.
200. McCabe, M.T., E.K. Lee, and P.M. Vertino, *A multifactorial signature of DNA sequence and polycomb binding predicts aberrant CpG island methylation.* Cancer Res, 2009. **69**(1): p. 282-91.
201. Saxonov, S., P. Berg, and D.L. Brutlag, *A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.* Proc Natl Acad Sci U S A, 2006. **103**(5): p. 1412-7.
202. Pasini, D., et al., *Regulation of stem cell differentiation by histone methyltransferases and demethylases.* Cold Spring Harb Symp Quant Biol, 2008. **73**: p. 253-63.
203. Hanahan, D. and L.M. Coussens, *Accessories to the crime: functions of cells recruited to the tumor microenvironment.* Cancer Cell, 2012. **21**(3): p. 309-22.
204. Yousfi, M., F. Lasmoles, and P.J. Marie, *TWIST inactivation reduces CBFA1/RUNX2 expression and DNA binding to the osteocalcin promoter in osteoblasts.* Biochem Biophys Res Commun, 2002. **297**(3): p. 641-4.
205. Colla, S., et al., *Human myeloma cells express the bone regulating gene Runx2/Cbfa1 and produce osteopontin that is involved in angiogenesis in multiple myeloma patients.* Leukemia, 2005. **19**(12): p. 2166-76.
206. Fujii, S., et al., *Enhancer of zeste homologue 2 (EZH2) down-regulates RUNX3 by increasing histone H3 methylation.* J Biol Chem, 2008. **283**(25): p. 17324-32.
207. Liu, Y., Z. Shao, and G.-C. Yuan, *Prediction of Polycomb target genes in mouse embryonic stem cells.* Genomics, 2010. **96**(1): p. 17-26.